Applications of Big Data

Unit I

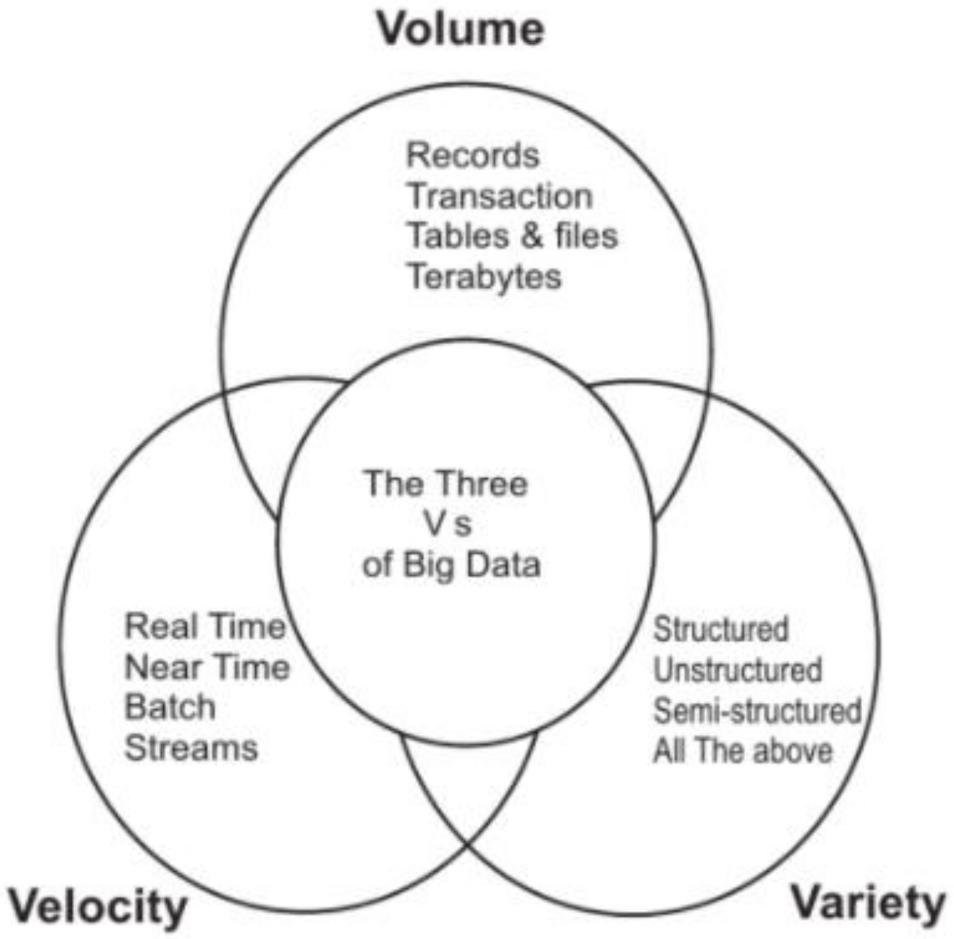
Introduction to Big Data, Characteristics of Data, and Big Data Evolution of Big Data, Definition of Big Data, Challenges with big data, Why Big data? Data Warehouse environment, Traditional Business Intelligence versus Big Data. State of Practice in Analytics, Key roles for New Big Data Ecosystems, Examples of Big Data Analytics.

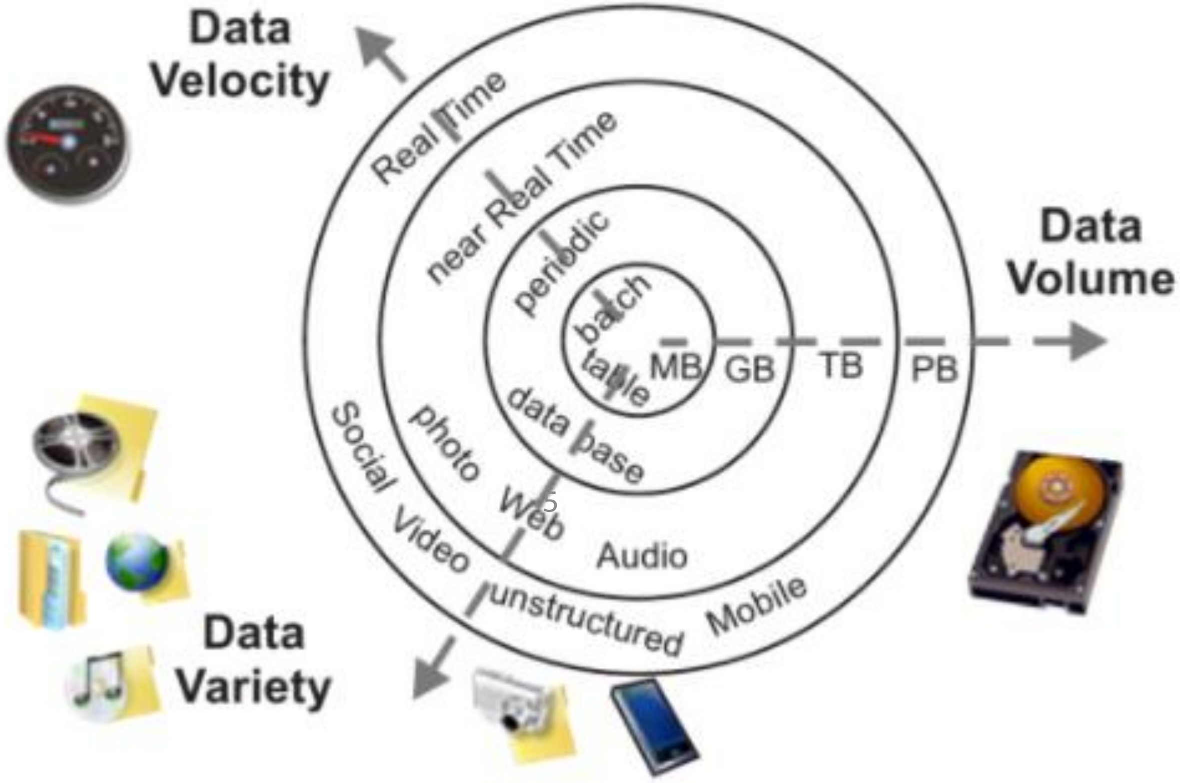
Big Data Analytics, Introduction to big data analytics, Classification of Analytics, Challenges of Big Data, Importance of Big Data, Big Data Technologies, Data Science, Responsibilities, Soft state eventual consistency. Data Analytics Life Cycle.

Topics

- Big Data Overview
  - What is Big Data?
  - Data -> Big Data
  - Definition of Big Data
  - Data Structure / Types of Big Data
  - Growth of Data - Tools and Techniques
  - Characteristics of Data
  - Evolution of Big Data
  - Characteristics of Big Data
  - Challenges With Big Data
  - Advantages of Big Data
  - Disadvantages of Big Data
- Applications of Big Data
- Why Big Data
- BI vs Big Data
- DW Environment
- Hadoop Environment
- Coexistence of Big Data & DW
- Analysts Perspective on Big Data
- State of the Practice in Analytics

V’s of Big Data





Volume (Size of Data)

- The name Big Data itself is related to a size which is enormous.
- The size of data plays a very crucial role in determining value out of data.
- Volume refers to the amount of data that exists.
- Volume is like the base of big data, as it is the initial size and amount of data that is collected.
- If the volume of data is large enough, it can be considered big data.
- What is considered to be big data is relative, though, and will change depending on the available computing power that's on the market.

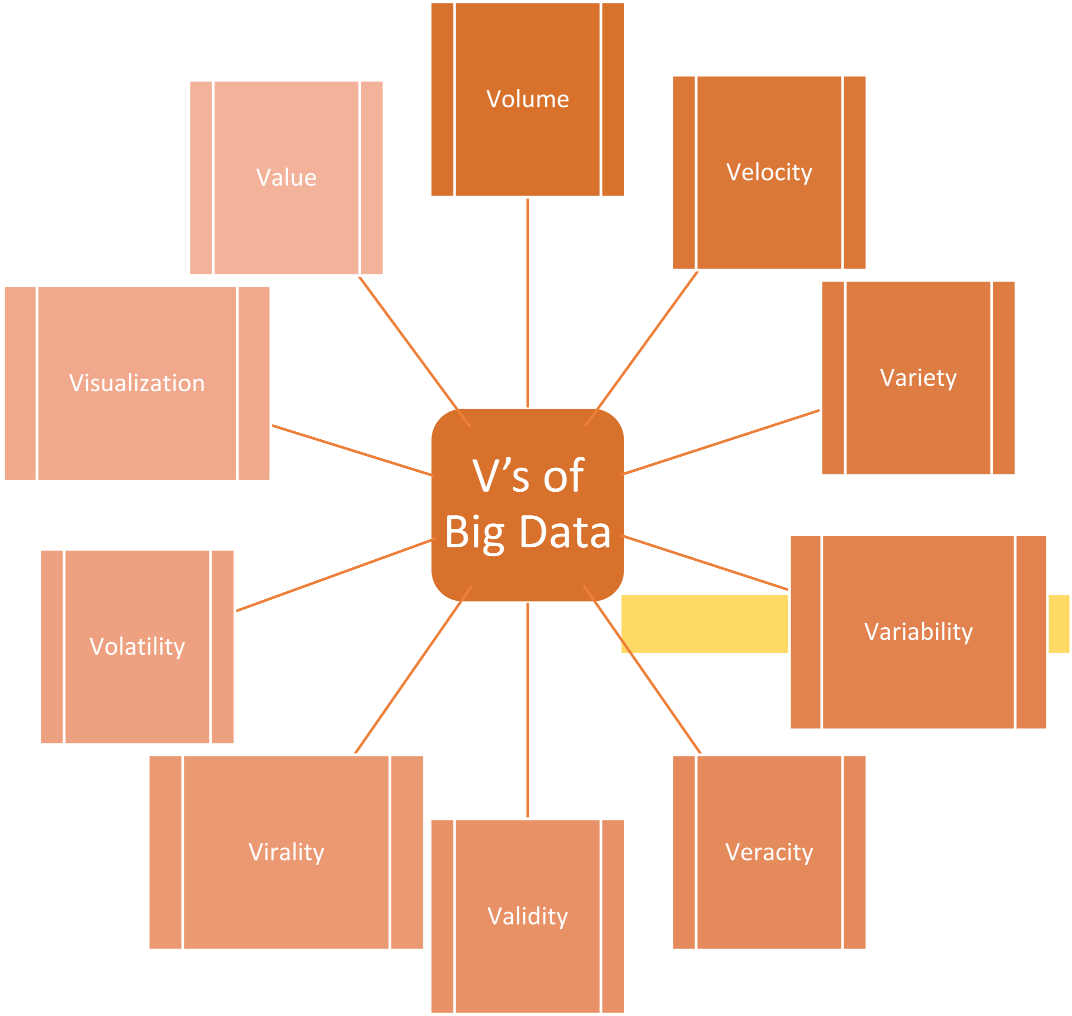
Velocity (Speed of Data)

- The term Velocity refers to the speed of generation of data, producing data, creating, refreshing data and moving data.
- This is an important aspect for companies need that need their data to flow quickly, so it's available at the right times to make the best business decisions possible.
- An organization that uses big data will have a large and continuous flow of data that is being created and sent to its end destination.
- Data could flow from sources such as machines, smartphones or social media, business processes, application logs, networks and social media sites, sensors. This data needs to be digested and analyzed quickly, and sometimes in near real time
- As an example, in healthcare, there are many medical devices made today to monitor patients and collect data. From in-hospital medical equipment to wearable devices, collected data needs to be sent to its destination and analyzed quickly.

Variety (Type of Data)

- Data comes in all types of formats – from structured datasets, numeric data in traditional databases to unstructured text documents, email, video, audio, stock ticker data and financial transactions.
- Variety refers to heterogeneous sources and the nature of data, structured, semi-structured and unstructured data as well.
- During earlier days, spreadsheets and databases were the only sources of data considered by most of the applications.
- Now days, data in the form of emails, photos, videos, monitoring devices, PDFs, audio, etc. is also being considered in the analysis applications.
- This variety of unstructured data poses certain issues for storage, mining and analysing data.
- The challenge in variety concerns the standardization and distribution of all data being collected.

Other V’s of Big Data



Variability(Data Differentiation)

- Data arrives constantly from different sources and how efficiently it differentiates between noisy data or important data
- Variability in big data's context refers to a few different things.
- One is the number of inconsistencies in the data, as the data is sourced from different sources.
- These need to be found by anomaly and outlier detection methods in order for any meaningful analytics to occur.
- Big data is also variable because of the multitude of data dimensions resulting from multiple disparate data types and sources.
- Variability can also refer to the inconsistent speed at which big data is loaded into your database.

Veracity (Quality of Data)

- Both value and veracity help define the quality and insights gathered from data. These particular characteristics also help determine whether the data is coming from a reliable source or the right fit for the analytic model.
- Accurate analysis of captured data is virtually worthless if it's not accurate
- It refers to the quality and accuracy of data.

  Gathered data could have missing pieces, may be inaccurate or may not be able to provide real, valuable insight. Veracity, overall, refers to the level of trust there is in the collected data.
- Data can sometimes become messy and difficult to use.

  A large amount of data can cause more confusion than insights if it's incomplete. For example, concerning the medical field, if data about what drugs a patient is taking is incomplete, then the patient's life may be endangered.

Validity (Data Authancity)

- Similar to veracity, validity refers to how accurate and correct the data is for its intended use.

- Correctness or accuracy of data used to extract result in the form of information.

- As the meaning of the word suggests, the validity of big data means how correct the data is for its purpose.

- Interestingly a considerable portion of big data remains un-useful, which is considered as '*dark data*.' The remaining part of collected unstructured data is cleansed first for analysis.

Virality (Spreading Speed)

- It is defined as the rate at which the data is broadcast /spread by a user and received by different users for their use

Volatility (Duration of Usefulness)

- Big data volatility means the stored data and how long is useful to the user.
- How old does your data need to be before it is considered irrelevant, historic, or not useful any longer? How long does data need to be kept for?
- Before big data, organizations tended to store data indefinitely -- a few terabytes of data might not create high storage expenses; it could even be kept in the live database without causing performance issues.
- Due to the velocity and volume of big data, however, its volatility needs to be carefully considered. You now need to establish rules for data currency and availability as well as ensure rapid retrieval of information when required.
- Make sure these are clearly tied to your business needs and processes -- with big data the costs and complexity of a storage and retrieval process are magnified.

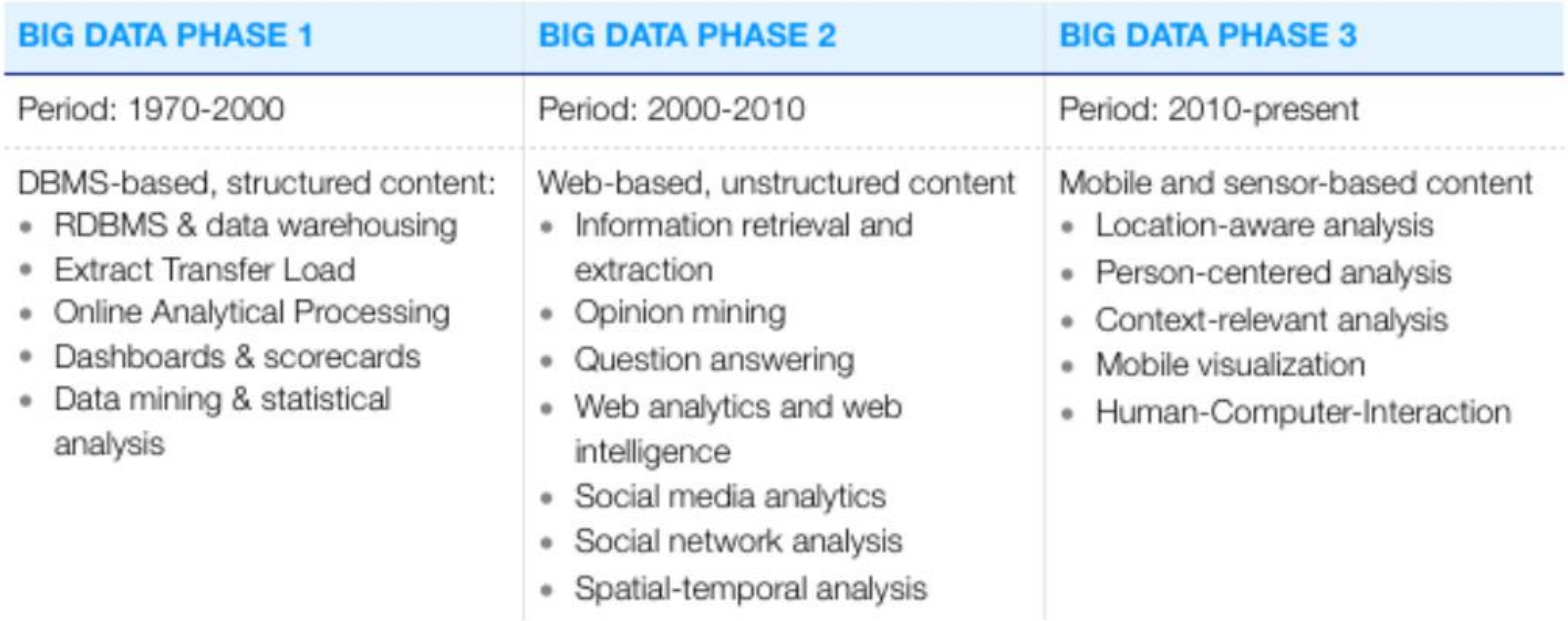
Visualization (Data Act/ Data Process)

- It is a process of representing abstract
- Big data processing is not the only means of getting a meaningful result out of it. Unless it is represented or visualizes in a meaningful way, there is no point in analyzing it.
- Hence, big data must be visualized with appropriate tools that serve different parameters to help data scientists or analysts understand it better.
- However, plotting billions of data points is not an easy task. Furthermore, it associates different techniques like using treemaps, network diagrams, cone trees, etc.

Value (Importance of Data)

- It represents the business value to be derived from big data
- This refers to the value that big data can provide, and it relates directly to what organizations can do with that collected data.
- Being able to pull value from big data is a requirement, as the value of big data increases significantly depending on the insights that can be gained from them.
- Substantial value can be found in big data, including understanding your customers better, targeting them accordingly, optimizing processes, and improving machine or business performance.
- You need to understand the potential, along with the more challenging characteristics, before embarking on a big data strategy.

History of Big Data

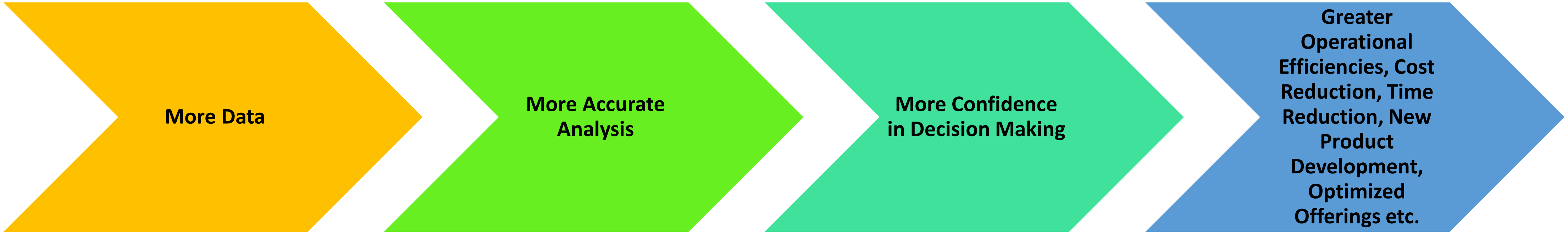


BIG DATA PHASE 1	BIG DATA PHASE 2	BIG DATA PHASE 3
Period: 1970-2000	Period: 2000-2010	Period: 2010-present
DBMS-based, structured content: • RDBMS & data warehousing • Extract Transfer Load • Online Analytical Processing • Dashboards & scorecards • Data mining & statistical analysis	Web-based, unstructured content • Information retrieval and extraction • Opinion mining • Question answering • Web analytics and web intelligence • Social media analytics • Social network analysis • Spatial-temporal analysis	Mobile and sensor-based content • Location-aware analysis • Person-centered analysis • Context-relevant analysis • Mobile visualization • Human-Computer-Interaction

Big Data Applications

- Retail
- Health care providers
- Education
- E-Commerce
- Media and Entertainment
- Finance
- Travel Industry
- Telecom
- Automobile
- Government Sector
- Weather Pattern
- Banking

Why Big data?



More Data

More Accurate Analysis

More Confidence in Decision Making

Greater Operational Efficiencies, Cost Reduction, Time Reduction, New Product Development, Optimized Offerings etc.

Enhancing operational Efficiency, Reducing cost & time, innovating new products & services and optimizing existing services.

- Results in +ve impact on

BI vs Big Data

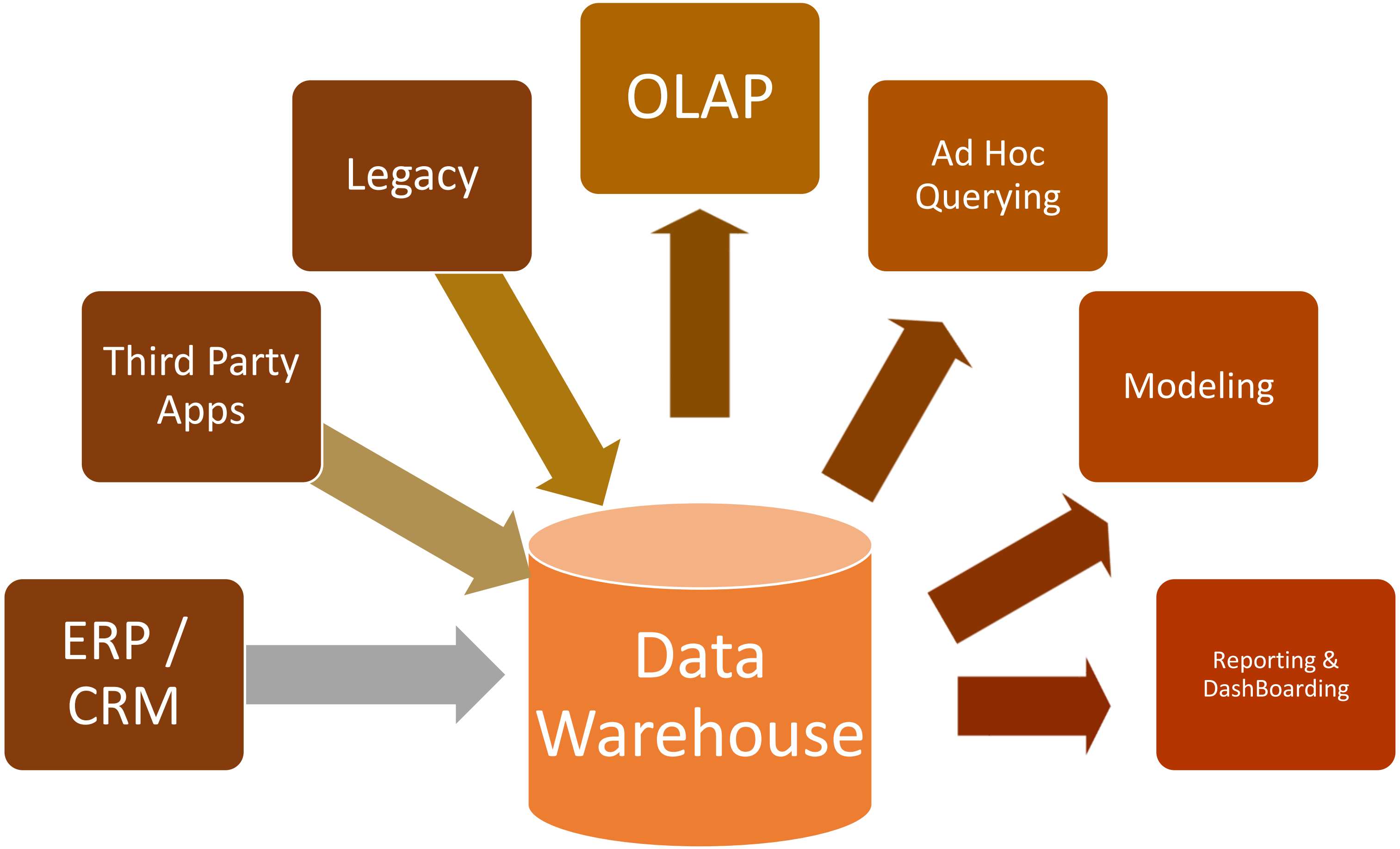
BI

- The Enterprise's Data is housed in Central Server
- Data is analysed in offline mode
- Traditional BI is about only Structured data

Big Data

- Big Data Environment Data resides in a Distributd File System
- Data is analysed in offline as well as real time mode
- Big data is about veriety of data

DW Environment



Legacy

OLAP

Ad Hoc Querying

Third Party Apps

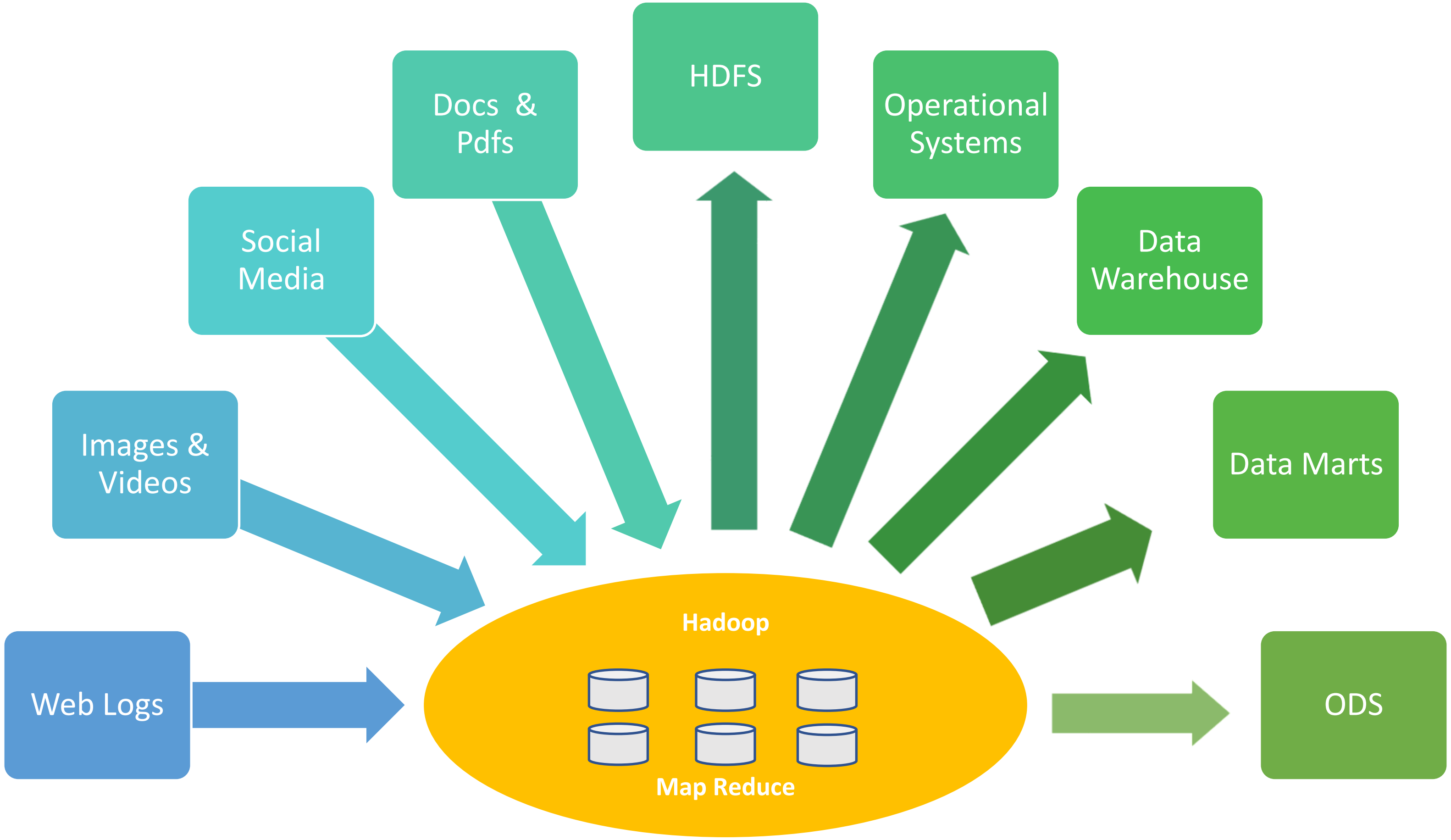
Modeling

ERP / CRM

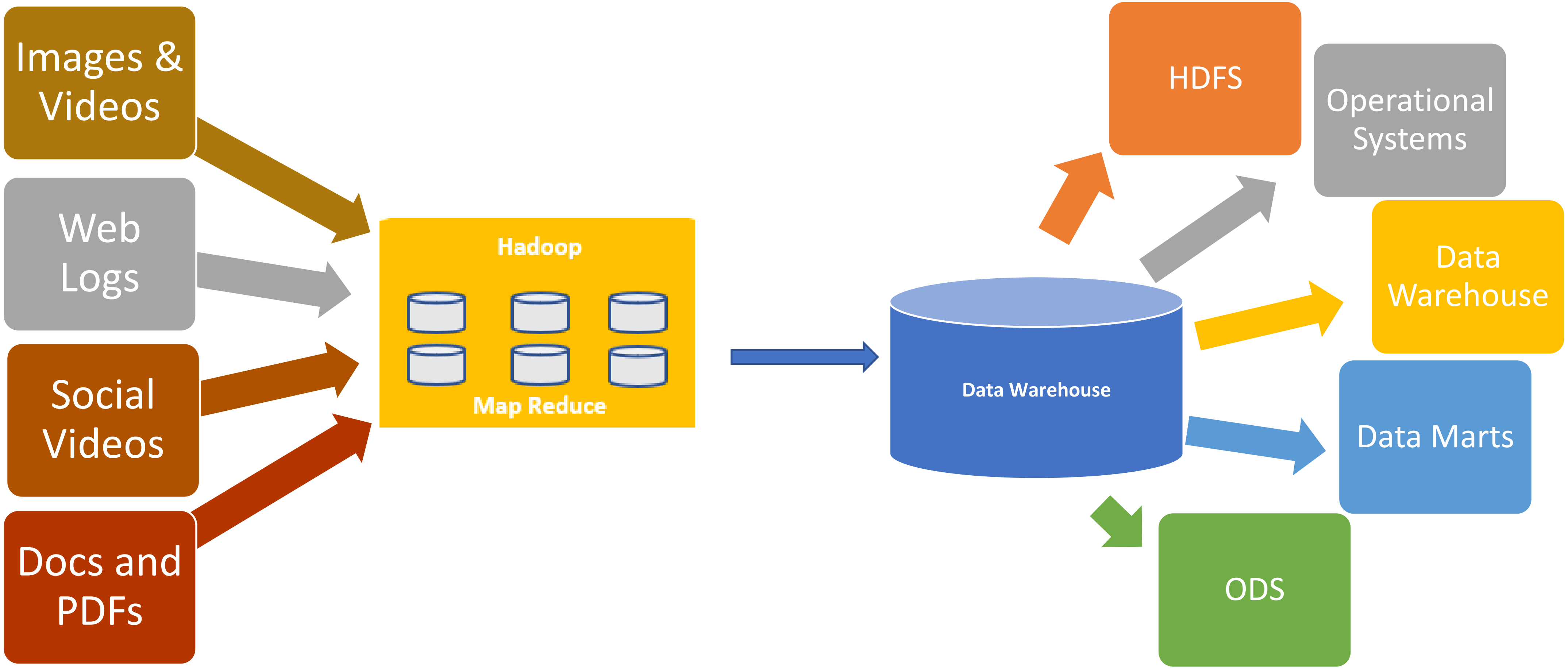
Data Warehouse

Reporting & DashBoarding

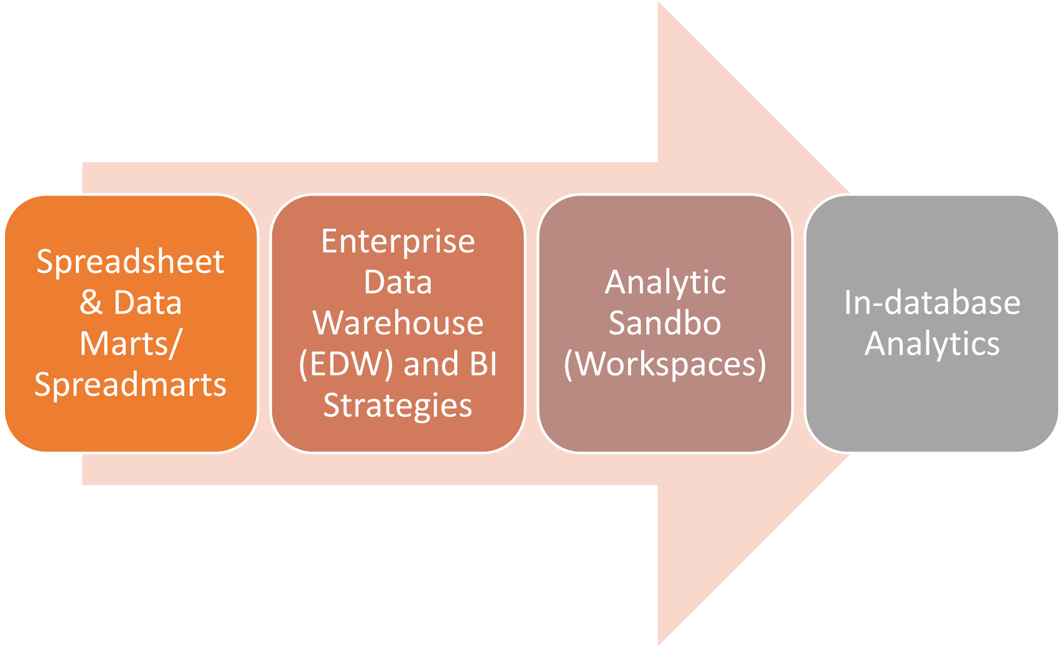
Hadoop Environment



Coexistence of Big Data & DW



Analyst Perspective on Big Data


Spreadsheet & Data Marts/ Spreadmarts
Enterprise Data Warehouse (EDW) and BI Strategies
Analytic Sandbo (Workspaces)
In-database Analytics

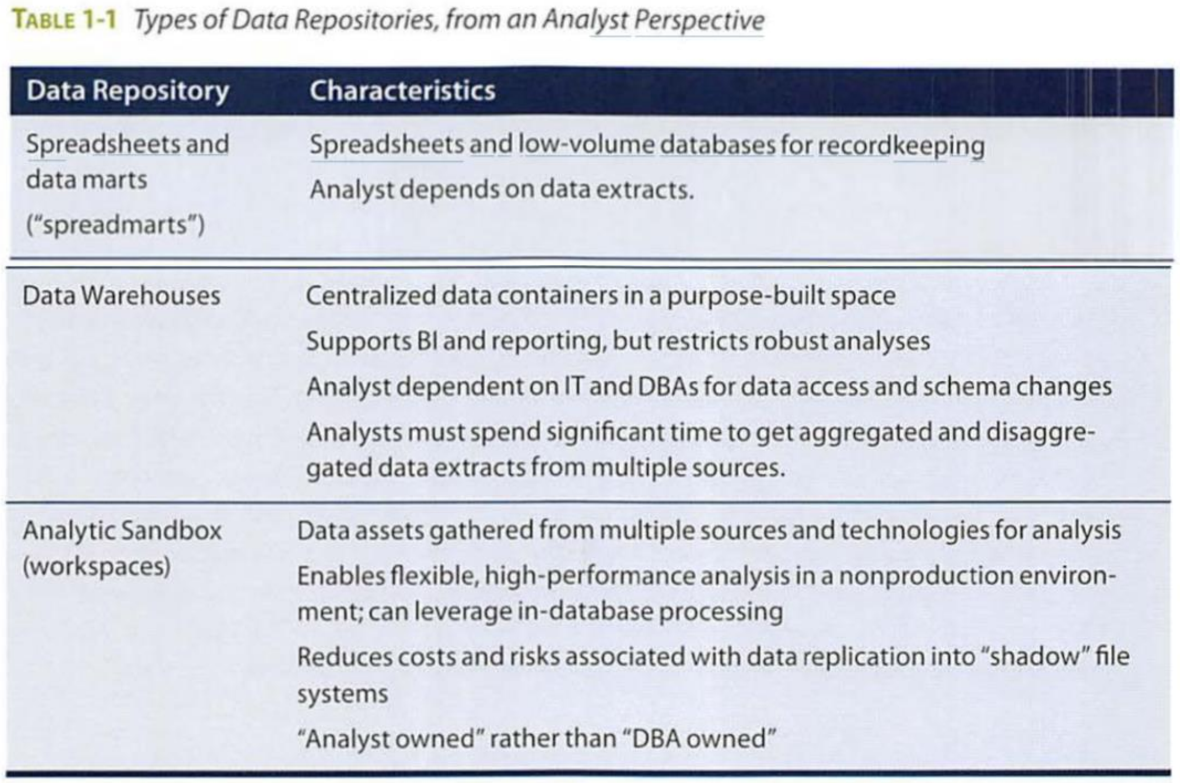


TABLE 1-1 *Types of Data Repositories, from an Analyst Perspective*

Data Repository	Characteristics
Spreadsheets and data marts ("spreadmarts")	Spreadsheets and low-volume databases for recordkeeping Analyst depends on data extracts.
Data Warehouses	Centralized data containers in a purpose-built space Supports BI and reporting, but restricts robust analyses Analyst dependent on IT and DBAs for data access and schema changes Analysts must spend significant time to get aggregated and disaggregated data extracts from multiple sources.
Analytic Sandbox (workspaces)	Data assets gathered from multiple sources and technologies for analysis Enables flexible, high-performance analysis in a nonproduction environment; can leverage in-database processing Reduces costs and risks associated with data replication into "shadow" file systems "Analyst owned" rather than "DBA owned"

State of the Practice in Analytics

State of Practice in Analytics

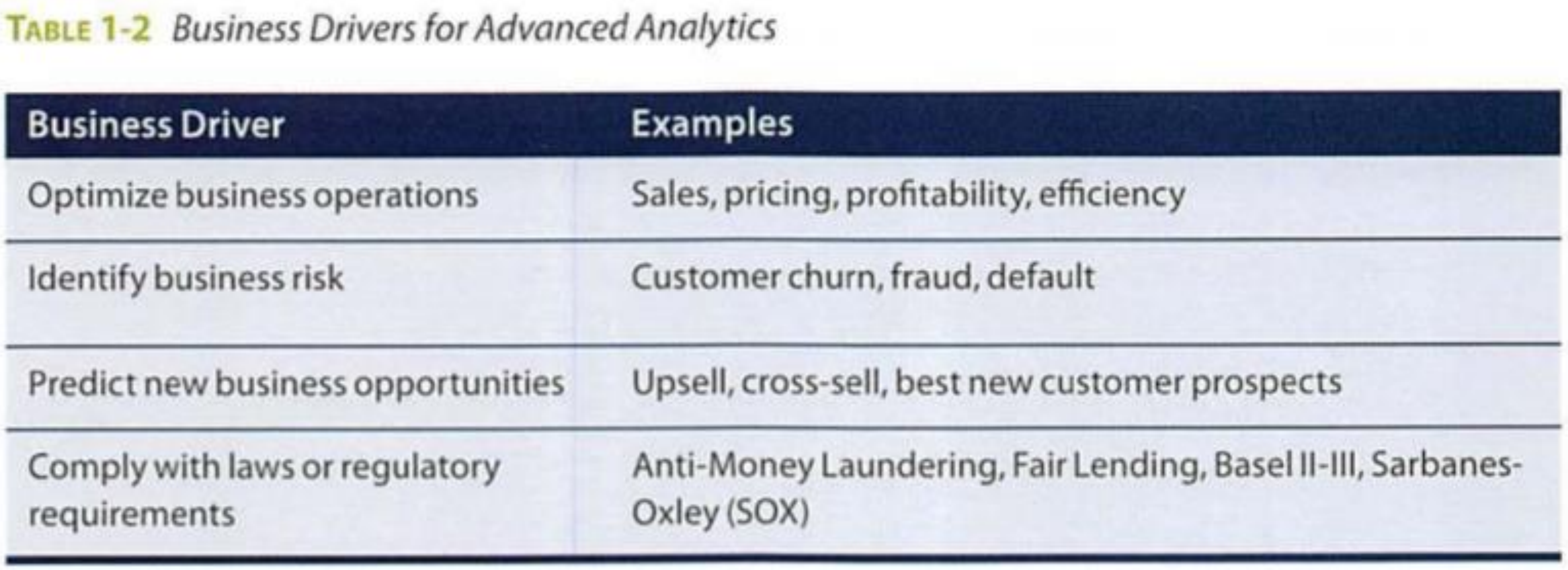
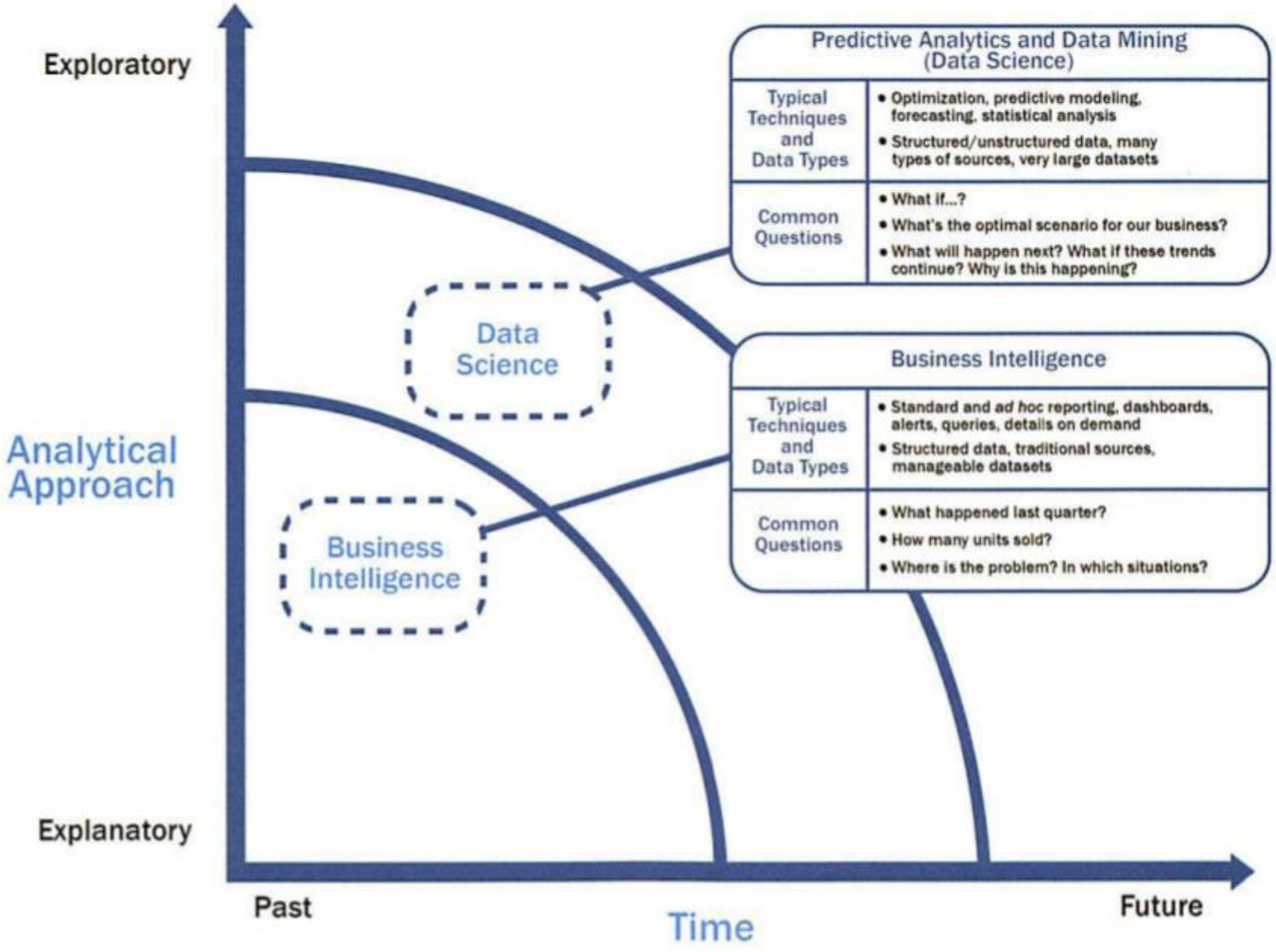


TABLE 1-2 *Business Drivers for Advanced Analytics*

Business Driver	Examples
Optimize business operations	Sales, pricing, profitability, efficiency
Identify business risk	Customer churn, fraud, default
Predict new business opportunities	Upsell, cross-sell, best new customer prospects
Comply with laws or regulatory requirements	Anti-Money Laundering, Fair Lending, Basel II-III, Sarbanes-Oxley (SOX)

BI vs DS



Predictive Analytics and Data Mining (Data Science)	
Typical Techniques and Data Types	• Optimization, predictive modeling, forecasting, statistical analysis • Structured/unstructured data, many types of sources, very large datasets
Common Questions	• What if...? • What's the optimal scenario for our business? • What will happen next? What if these trends continue? Why is this happening?

Business Intelligence	
Typical Techniques and Data Types	• Standard and *ad hoc* reporting, dashboards, alerts, queries, details on demand • Structured data, traditional sources, manageable datasets
Common Questions	• What happened last quarter? • How many units sold? • Where is the problem? In which situations?

FIGURE 1-8 *Comparing BI with Data Science*

Current Analytical Architecture

- 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
- 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
- Once in the data warehouse, data is read by additional applications across the enterprise for Bl and reporting purposes.
- At the end of this workflow, analysts get data provisioned for their downstream analytics.

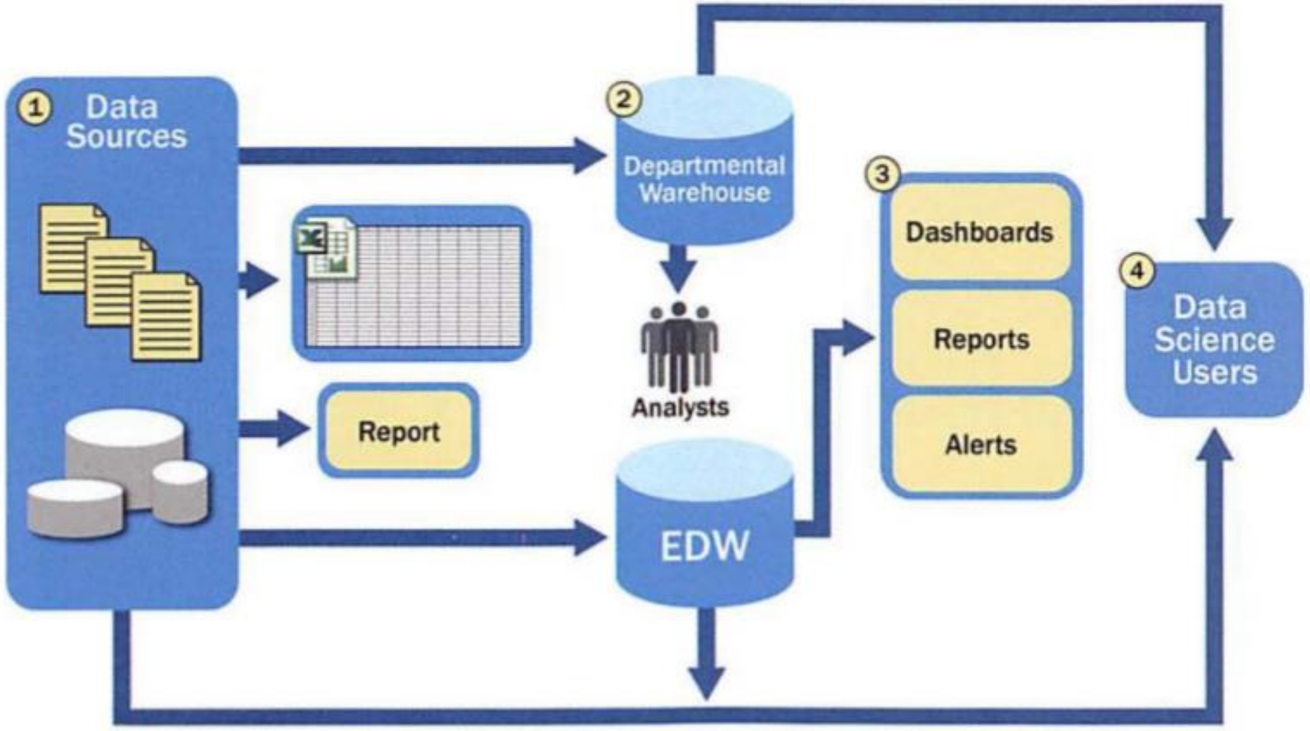
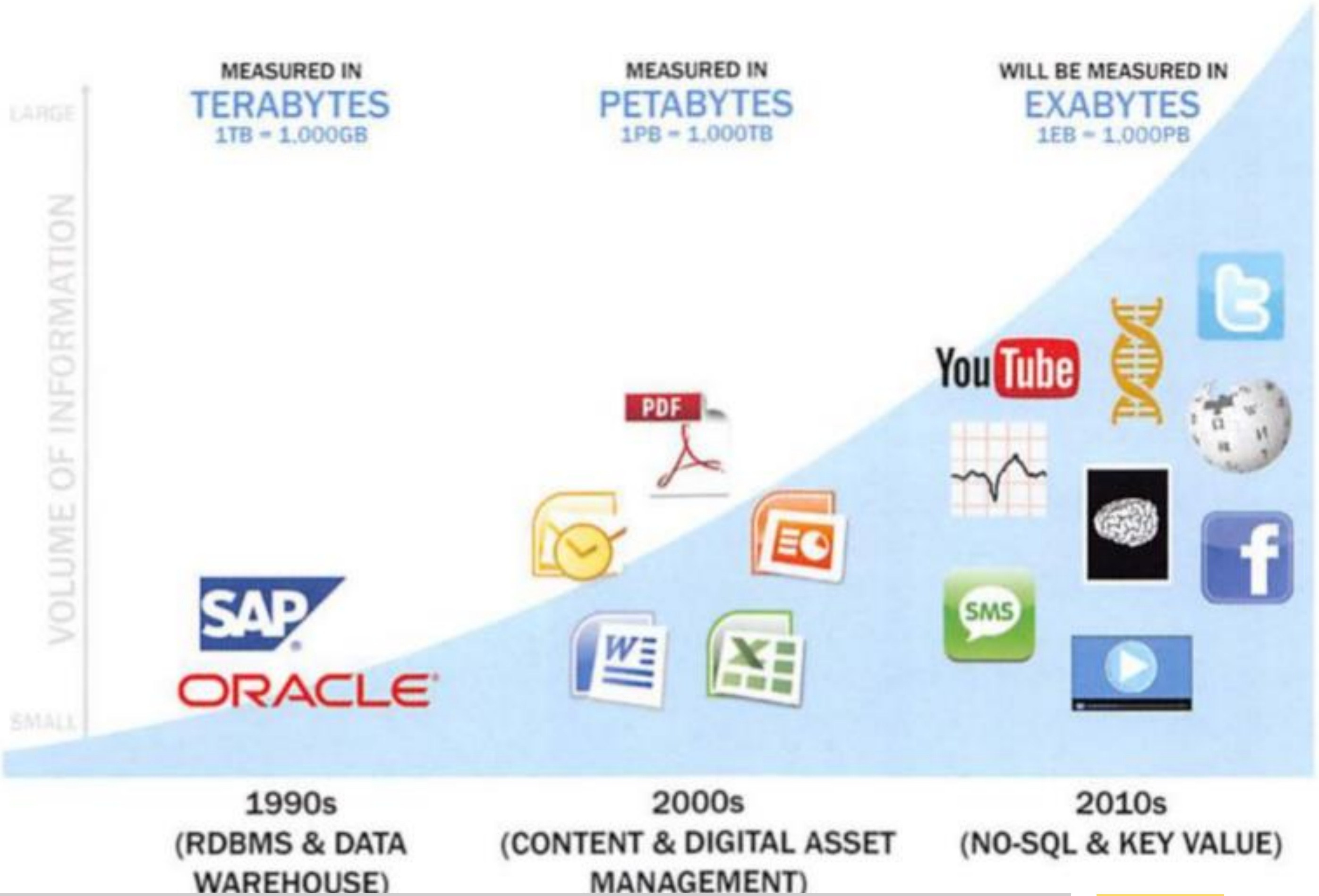


FIGURE 1-9 *Typical analytic architecture*

Drivers of Big Data

- Medical information, such as genomic sequencing and diagnostic imaging
- 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
- Nontraditional IT devices, including the use of radio-frequency identification (RFID) readers, GPS navigation systems, and seismic processing



Emerging Big Data Ecosystem & a New Approach to Analytics

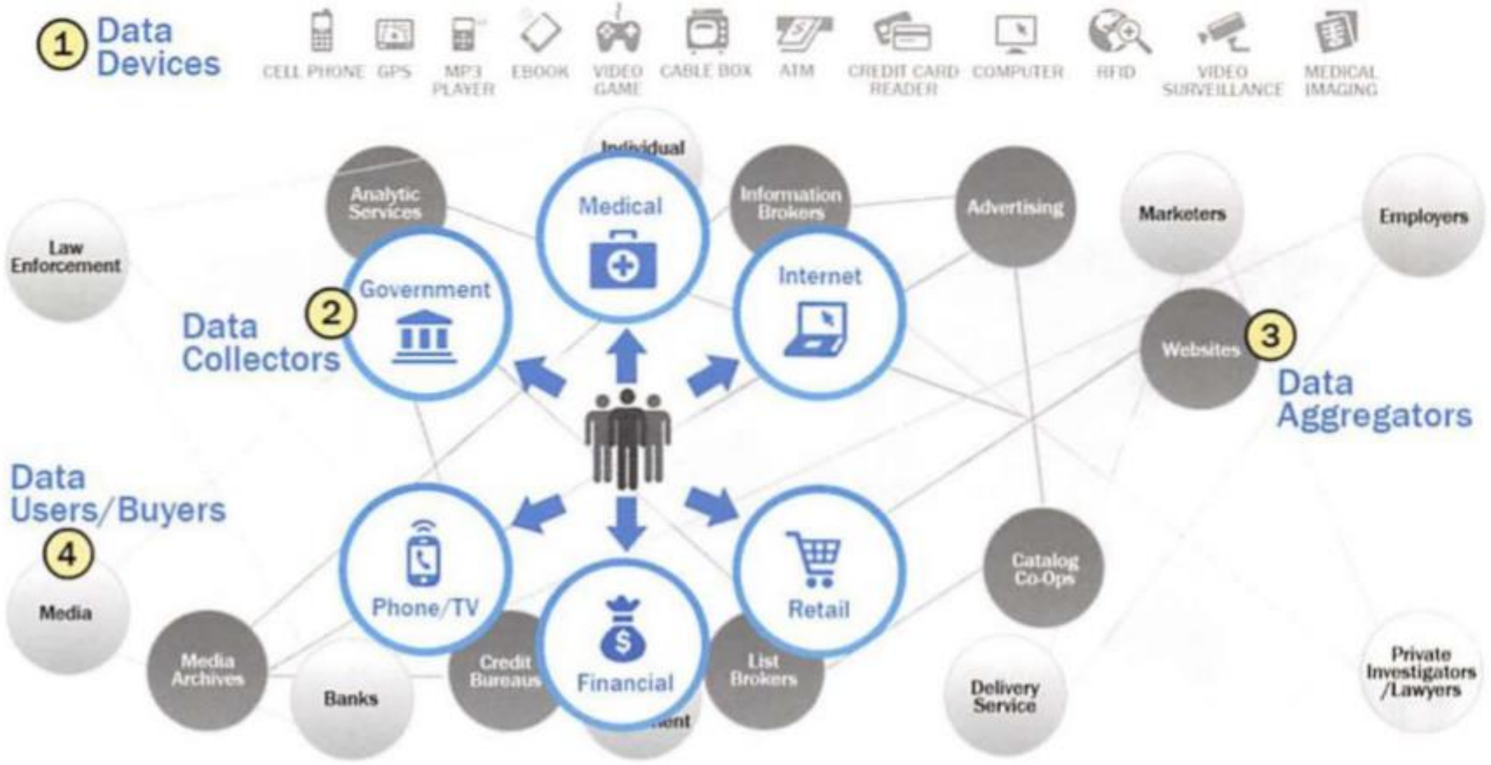
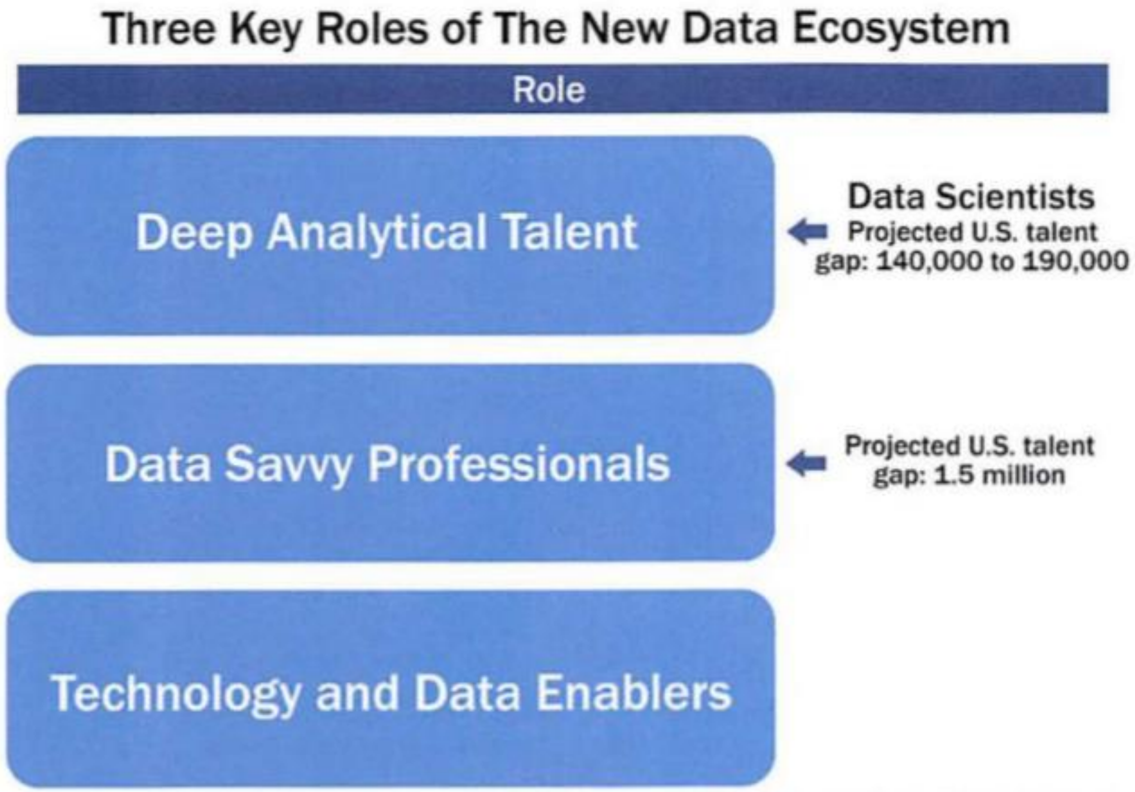


FIGURE 1-11 *Emerging Big Data ecosystem*

Key Roles for the Big Data Ecosystem



Three Key Roles of The New Data Ecosystem
Role
Deep Analytical Talent
Data Scientists
Projected U.S. talent gap: 140,000 to 190,000
Data Savvy Professionals
Projected U.S. talent gap: 1.5 million
Technology and Data Enablers

Note: Figures above reflect a projected talent gap in US in 2018, as shown in McKinsey May 2011 article "Big Data: The Next Frontier for Innovation, Competition, and Productivity"

FIGURE 1-12 *Key roles of the new Big Data ecosystem*

Data Scientist – Skill Set

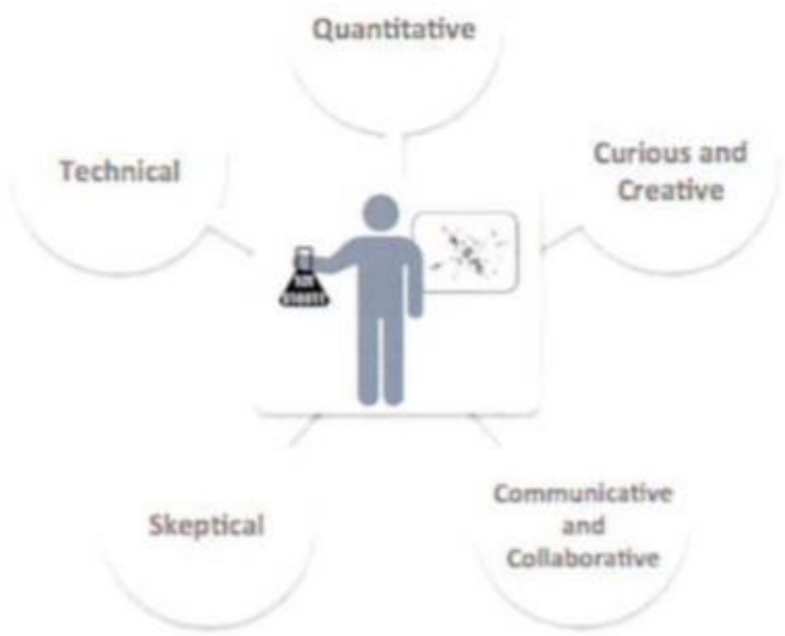


FIGURE 1-13 *Profile of a Data Scientist*

Thank You….

Revise the topics from Syllabus References…

Fill Your Attendance Form….!