Big Data Analytics

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 Covered

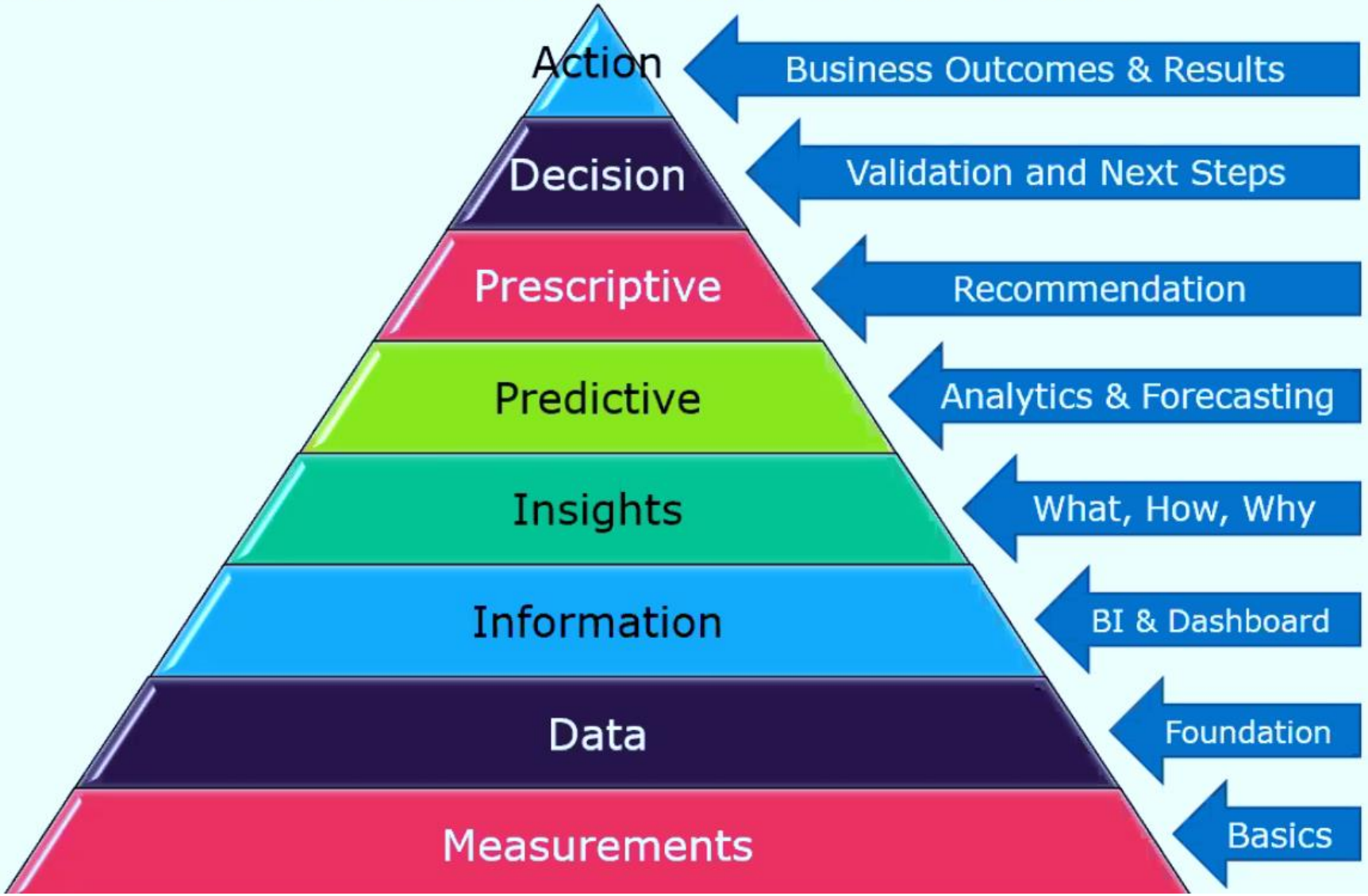
- Big Data Overview
  - What is Big Data?
  - Data -> Big Data
  - Definition of Big Data
  - Data Structure / Types of Big Data
  - Growth of Data - Tools and Techniques
  - Characteristics of Data
  - Evolution of Big Data
  - Characteristics of Big Data
  - Challenges With Big Data
  - Advantages 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
- BI VS DS
- Current Analytical Architecture
- Drivers of Big Data
- Emerging New Big Data Ecosystem
- Key Roles for the new Big Data Ecosystem
- Data Scientist Skill Set
- Examples of Big Data Analytics
- Big Data Analytics
- Importance of Big Data Analytics
- What Big Data Analytics is not?
- Classification of Big Data Analytics
- Big Data Technologies

Today's Topics

- Data Analytics Life Cycle
- Key Roles and Responsibilities for a Successful Analytics Project
- Background & Overview of Data Analytics Lifecycle
- Key Outputs from a Successful Analytical Projects

Data Analytics Life Cycle

Data... to Decision... to Action...



25

Understanding Past, Present & Future

	Past	Present	Future
Information	What Happened ***Reporting***	What is happening now ***Alerts***	What will happen What can happen ***Extrapolation***
Insights	How did it happen ***Modeling & Analytics***	What's the next best action ***Recommendations***	What's the best / worst can happen ***Prediction***
Action	Why did this happen and how to prevent / repeat it ***Root Cause Analysis***	Is this happening as expected ***Verification & Validation***	How are we going to handle it, Next best action ***Prescriptive***

Big Data Analytics

- Big data analytics examines large amounts of data to uncover hidden patterns, correlations and other insights. With today's technology, it's possible to analyze your data and get answers from it almost immediately – an effort that's slower and less efficient with more traditional business intelligence solutions.

Importance of BDA

1. **Cost reduction.** Big data technologies such as Hadoop and cloud-based analytics bring significant cost advantages when it comes to storing large amounts of data – plus they can identify more efficient ways of doing business.
2. **Faster, better decision making.** With the speed of Hadoop and in-memory analytics, combined with the ability to analyze new sources of data, businesses are able to analyze information immediately – and make decisions based on what they've learned.
3. **New products and services.** With the ability to gauge customer needs and satisfaction through analytics comes the power to give customers what they want. Davenport points out that with big data analytics, more companies are creating new products to meet customers' needs.

What Big Data Analytics is & is not?

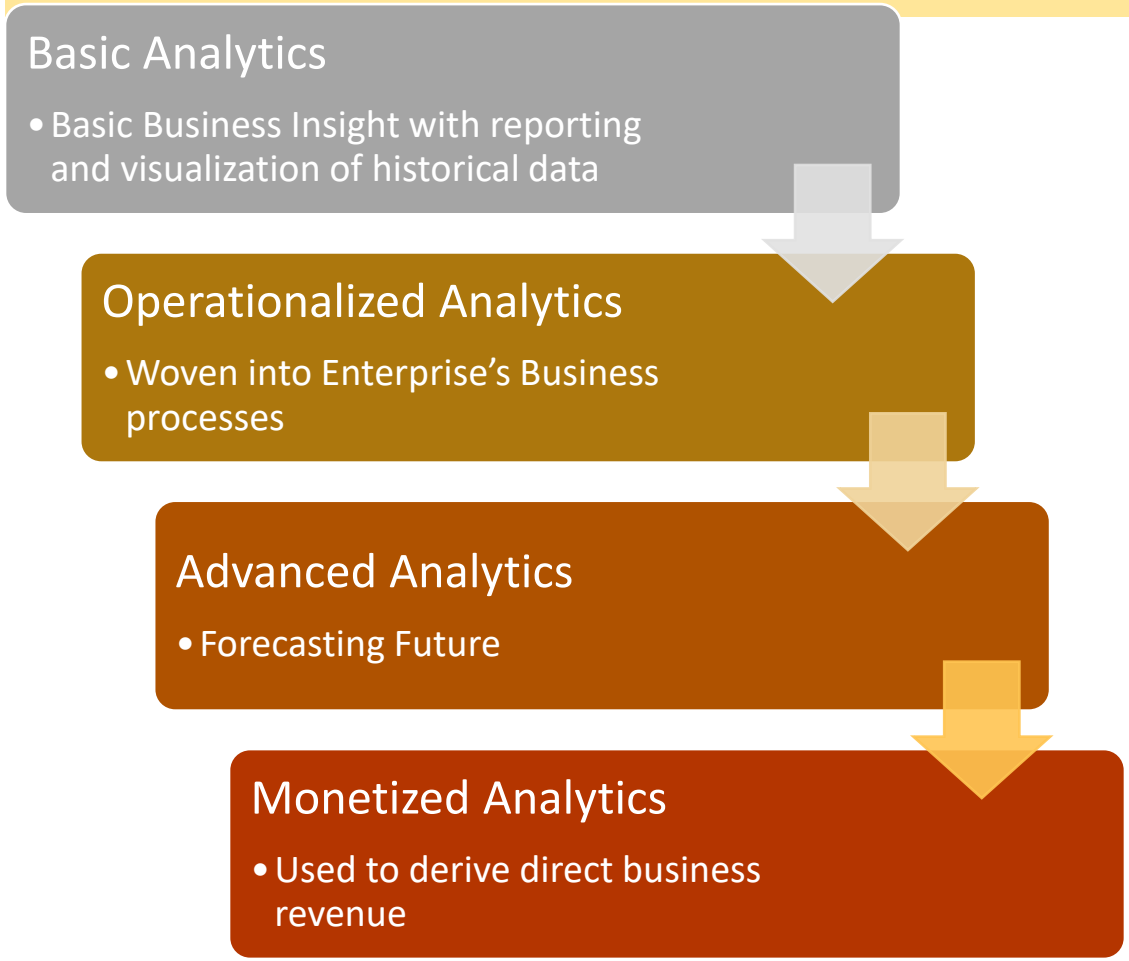
What is Big Data Analytics?

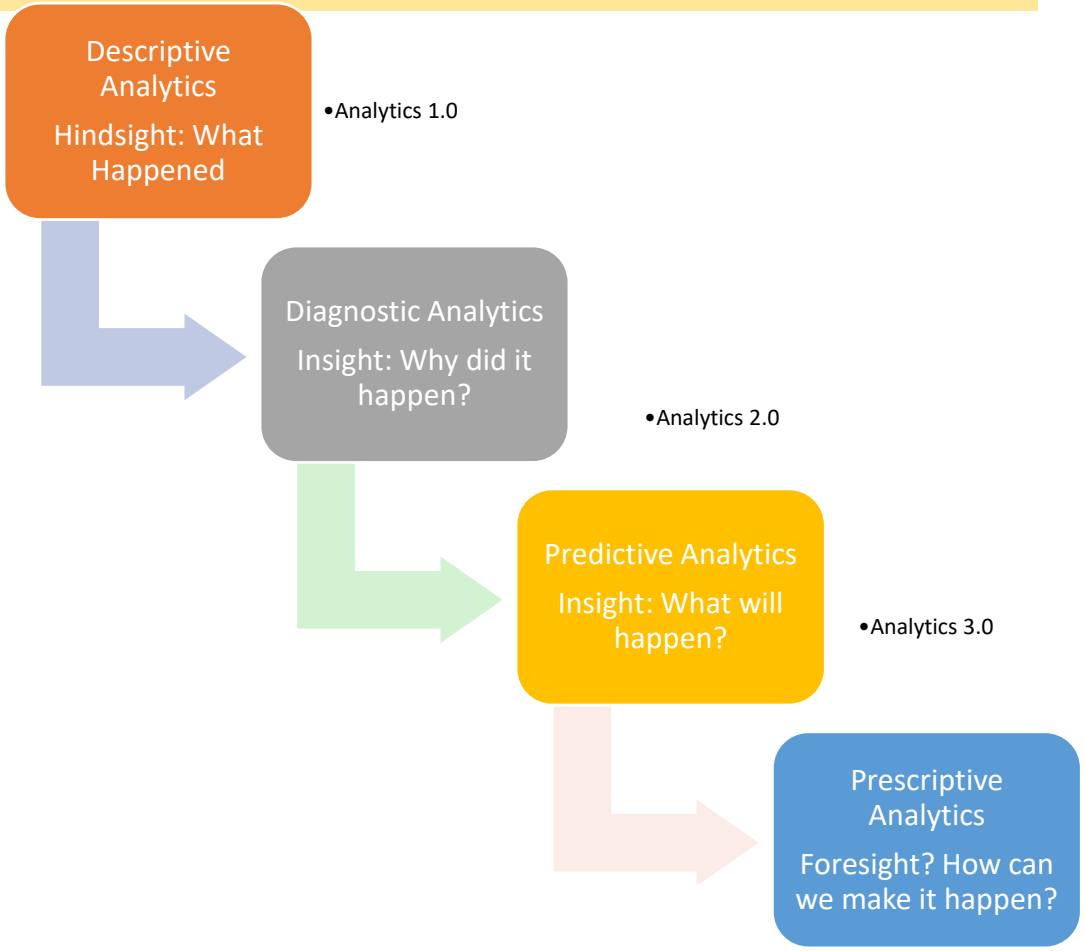
- Technology enabled Analytics
- Competitive advantage
- Reacher / Deeper business insight
- Real-Time Analytics
- Collaboration with data scientist and business users
- Working with huge datasets
- Better and faster decisions in real time
- More code to data with greater speed & efficiency

What Big Data Analytics is not?

- Only about Volume
- Just About Technology
- Meant to Replace RDBMS
- Meant to replace DW
- Only used by huge online companies like Google / Amazon
- One-size fill all RDBMS built on shared disk and memory

Classification of Big Data Analytics





17 February 2022

Phases of Analytics

Table 3.1 Analytics 1.0, 2.0, and 3.0

Analytics 1.0	Analytics 2.0	Analytics 3.0
Era: mid 1950s to 2009	2005 to 2012	2012 to present
Descriptive statistics (report on events, occurrences, etc. of the past)	Descriptive statistics + predictive statistics (use data from the past to make predictions for the future)	Descriptive + predictive + prescriptive statistics (use data from the past to make prophecies for the future and at the same time make recommendations to leverage the situation to one's advantage)
Key questions asked: What happened? Why did it happen?	Key questions asked: What will happen? Why will it happen?	Key questions asked: What will happen? When will it happen? Why will it happen? What should be the action taken to take advantage of what will happen
Data from legacy systems, ERP, CRM, and 3rd party applications.	Big data	A blend of big data and data from legacy systems, ERP, CRM, and 3rd party applications.
Small and structured data sources. Data stored in enterprise data warehouses or data marts.	Big data is being taken up seriously. Data is mainly unstructured, arriving at a much higher pace. This fast flow of data entailed that the influx of big volume data had to be stored and processed rapidly, often on massive parallel servers running Hadoop.	A blend of big data and traditional analytics to yield insights and offerings with speed and impact.
Data was internally sourced.	Data was often externally sourced.	Data is both being internally and externally sourced.
Relational databases	Database appliances, Hadoop clusters, SQL to Hadoop environments, etc.	In memory analytics, in database processing, agile analytical methods, machine learning techniques, etc.

Data Analytics Life Cycle

The Data Analytics Lifecycle is designed specifically for Big Data problems and data science projects. The lifecycle has six phases, and project work can occur in several phases at once.

Key Roles and Responsibilities for a successful Analytics Project

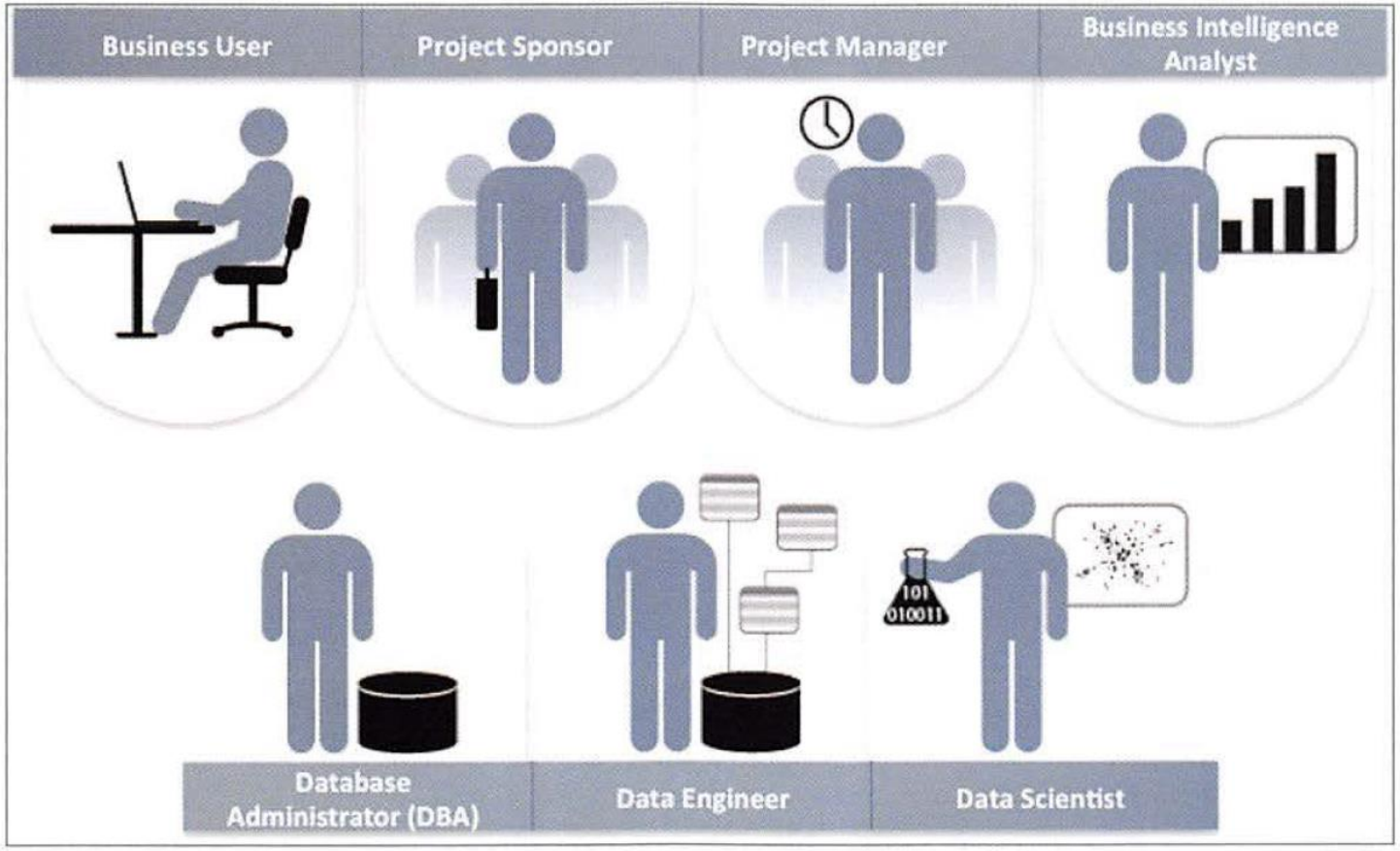
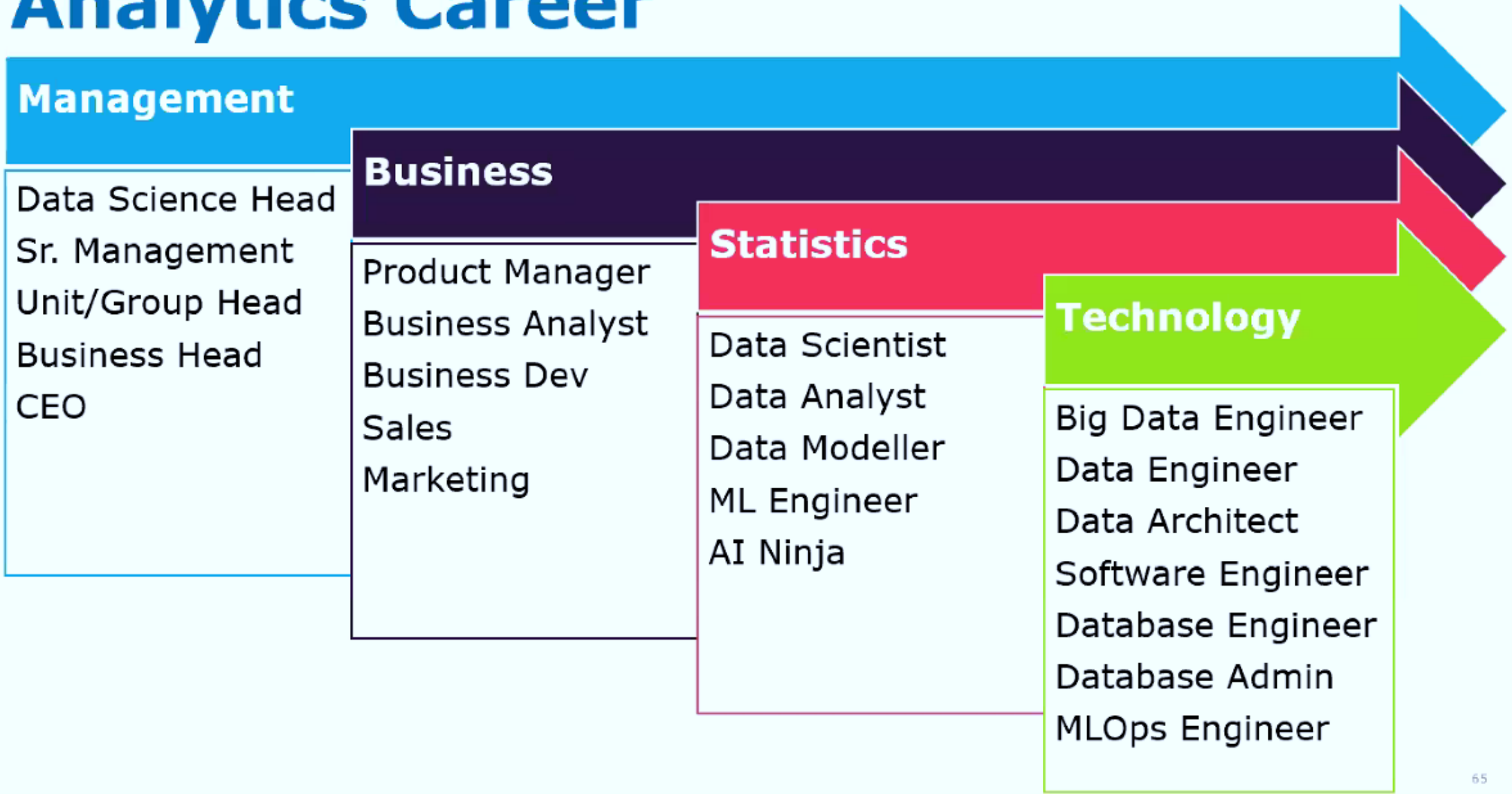


FIGURE 2-1 *Key roles for a successful analytics project*

Analytics Career



Background and Overview of Data Analytics Lifecycle

- Some Process

Scientific Method

CRISP-DM

Tom Davenports DELTA

Applied Information Economics

MAD Skills

Background and Overview of Data Analytics Lifecycle

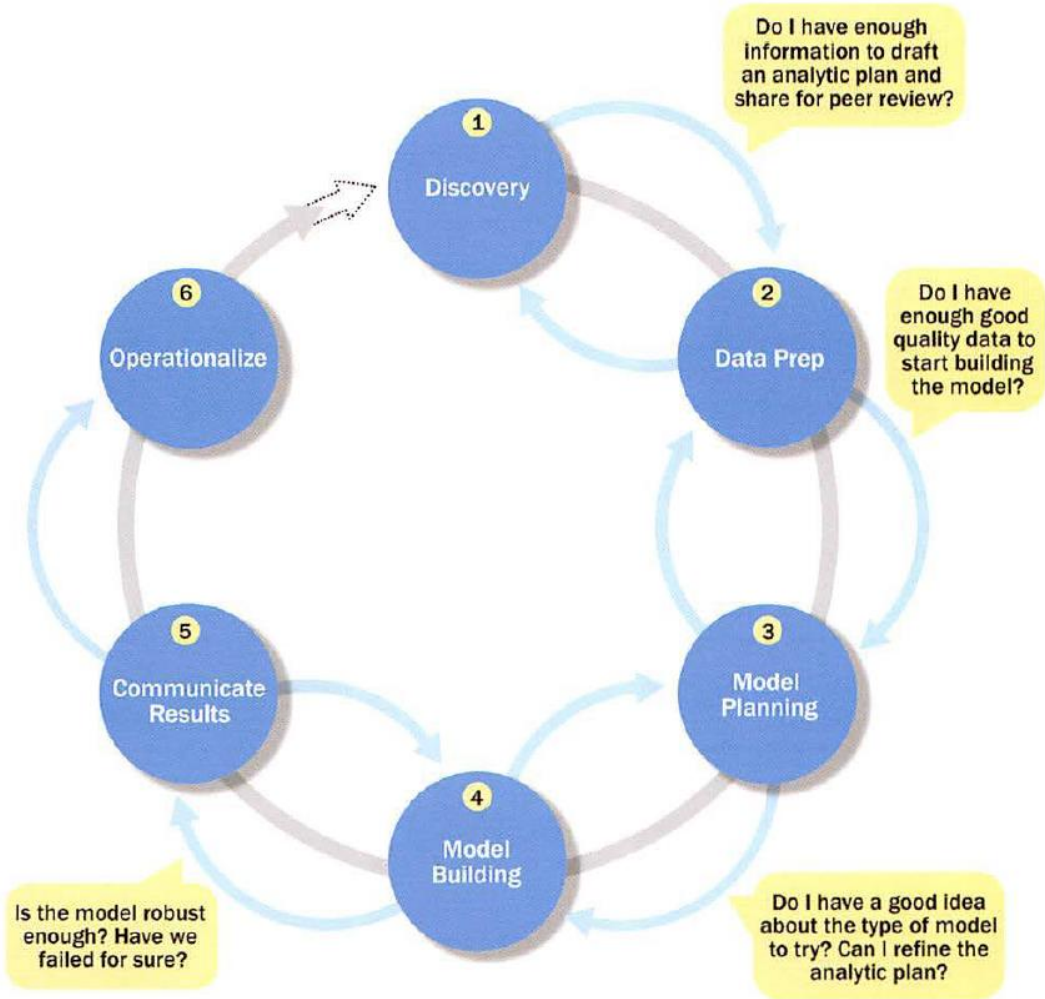


FIGURE 2-2 *Overview of Data Analytics Lifecycle*

Phase 1 Discovery

- Learning the Business Domain
- Resources
- Framing the Problem
- Identifying Key Stakeholders
- Interviewing the Analytics Sponsor
- Developing Intial Hypothesis
- Identifying Potential Data Sources
  - Identify Data Sources
  - Capture Aggregate Data Sources
  - Review the Raw Data
  - Evaluate the Data Structures and Tools Needed
  - Scoop the type of Data Infrastructures needed for this type of Problem

Phase 2 Data Preparation

- Preparing the Analytic Sandbox
- Performing ETLT
- Learning about the Data
- Data Conditioning
- Survey and Visualization
- Common Tool for Data Preparation Phase
  - Hadoop, Alpine Miner, OpenRefine, Data Wrangler

Phase 3 Model Planning

- Data Exploration and Variable Selection
- Model Selection
- Common Tools for Model Planning Phase
  - R
  - SQL Analysis Services
  - SAS/ACCESS with data connectors like ODBC, JDBC and OLEDB

Phase 4 Model Building

- Does the model appear valid and accurate on the test data?
- Does the model output/behavior make sense to the domain experts? That is, does it appear as if the model is giving answers that make sense in this context?
- Do the parameter values of the fitted model make sense in the context of the domain?
- Is the model sufficiently accurate to meet the goal?
- Does the model avoid intolerable mistakes? Depending on context, false positives may be more serious or less serious than false negatives, for instance.
- Are more data or more inputs needed? Do any of the inputs need to be transformed or eliminated?
- Will the kind of model chosen support the runtime requirements?
- Is a different form of the model required to address the business problem? If so, go back to the model planning phase and revise the modeling approach.

Phase 4 Model Building…

- Common Tools for Module Building
  - Commercial Tools
    - SAS Enterprise Miner
    - SPSS Modeler
    - MATLAB
    - Alpine Miner
    - Statistica & Mathematica
  - Open Source Tools
    - R, PL/R
    - Octave
    - WEKA
    - Python
    - SQL – In-Database Implications such as MADLib

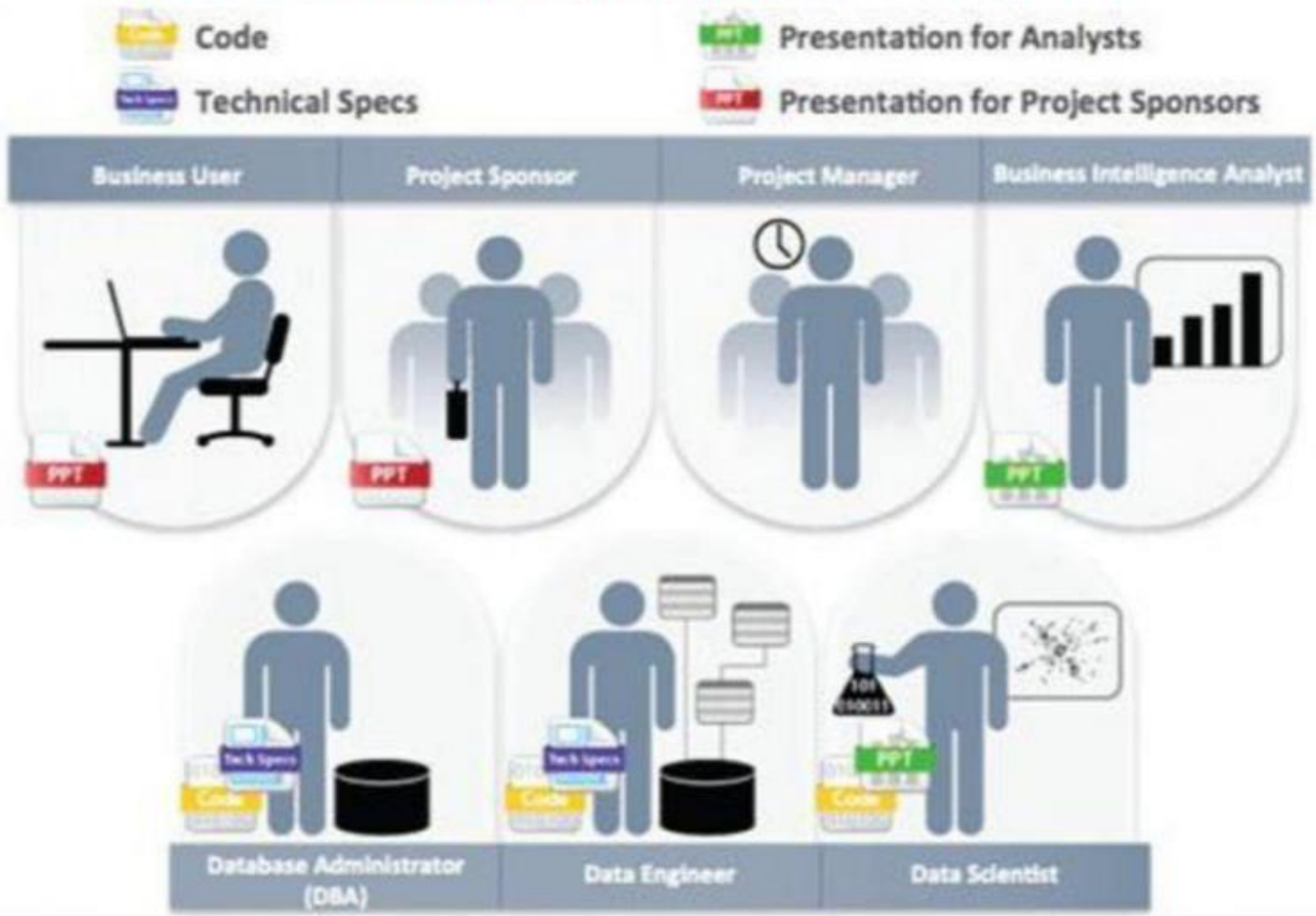
Phase 5 Communicate Results

- Sometimes teams have only done a superficial analysis, which is not robust enough to accept or reject a hypothesis.
- Other times, teams perform very robust analysis and are searching for ways to show results, even when results may not be there.
- It is important to strike a balance between these two extremes when it comes to analyzing data and being pragmatic in terms of showing real-world results.
- When conducting this assessment, determine if the results are statistically significant and valid.
- During this step, assess the results and identify which data points may have been surprising and which were in line with the hypotheses that were developed in Phase 1.
- As a result of this phase, the team will have documented the key findings and major insights derived from the analysis

Phase 6 Operationalize

- In the final phase, the team communicates the benefits of the project more broadly and sets up a pilot project to deploy the work in a controlled way
- While scoping the effort involved in conducting a pilot project, consider running the model in a production environment for a discrete set of products or a single line of business, which tests the model in a live setting.
- This allows the team to learn from the deployment and make any needed adjustments before launching the model across the enterprise.
- Part of the operationalizing phase includes creating a mechanism for performing ongoing monitoring of model accuracy and.
- if accuracy degrades, finding ways to retrain the model.
- Refer Fig. on next slide

Key Outputs from a Successful Analytics Project



Thank You….

Revise the topics from Syllabus References…

Fill Your Attendance Form….!

Syllabus References

1. Big Data and Analytics, Subhashini Chellappan Seema Acharya, Wiley

2. Data Analytics with Hadoop *An Introduction for Data Scientists*, Benjamin Bengfort and Jenny Kim, O'Reilly

3. Big Data and Hadoop, V.K Jain, Khanna Publishing

   https://books.google.co.in/books?id=i6NODQAAQBAJ&pg=PA122&source=gbs_toc_r&cad=4#v=onepage&q&f=true