SNS COLLEGE OF TECHNOLOGY

(An AUTONOMOUS INSTITUTION)

RE-ACCREDITED BY NAAC WITH A+ GRADE, ACCREDITED BY NBA(CSE, IT, ECE, EEE & MECHANICAL)

APPROVEDY BY AICTE, NEW DELHI, RECOGNIZED BY UGC, AFFILIATED TO ANNA UNIVERSITY, CHENNAI

TOPIC: ANALYTIC PROCESS AND TOOLS

COURSE: 19CAT702 Big Data Analytics

UNIT I: DATA - ANALYTIC PROCESSES AND TOOLS

ELECTIVE: III SEMESTER / II MCA

Session Objectives

- Demonstrate how data analytics works on various platforms
- Understand the role of data analytics in business
- Know the various tools available for data analytics

Data Analytics

- Process for obtaining raw data and converting it into information useful for decision-making
- Data is collected and analyzed to answer questions, test hypotheses
- Used to optimize business performance
- It is an intersection of information technology, statistics and business

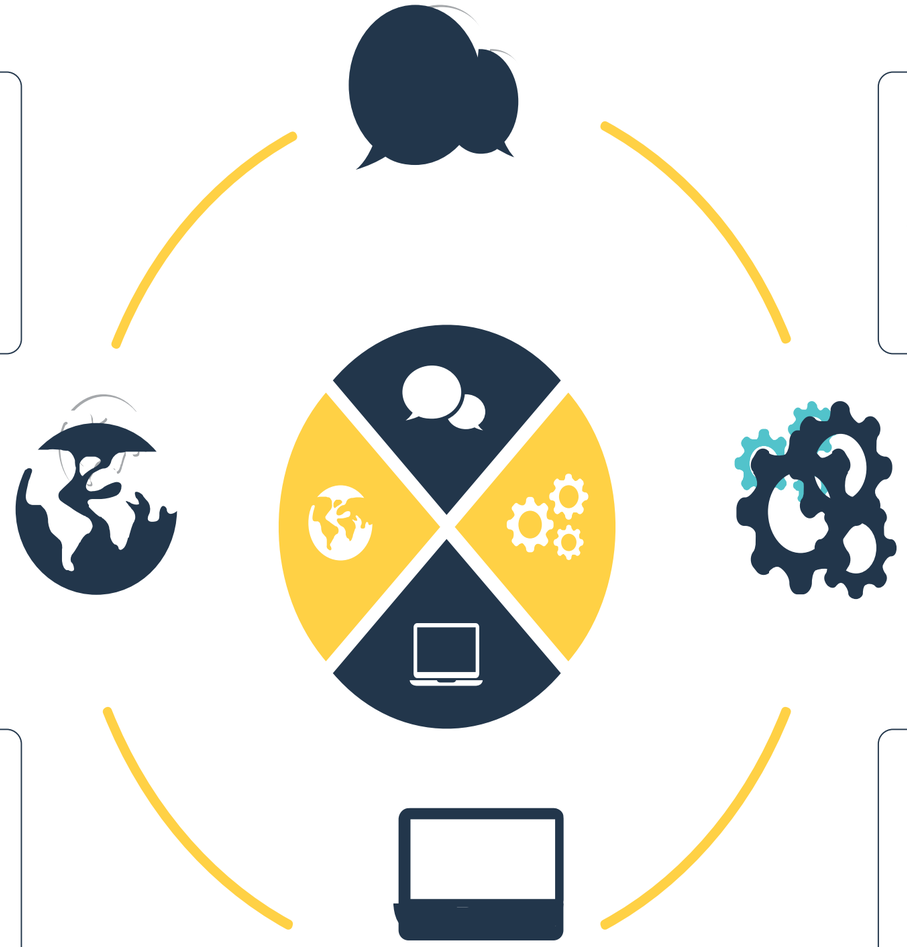
Data Analytics: Types

Descriptive analytics

what has happened over a given period of time?

Predictive Analysis

what is likely going to happen in the near term



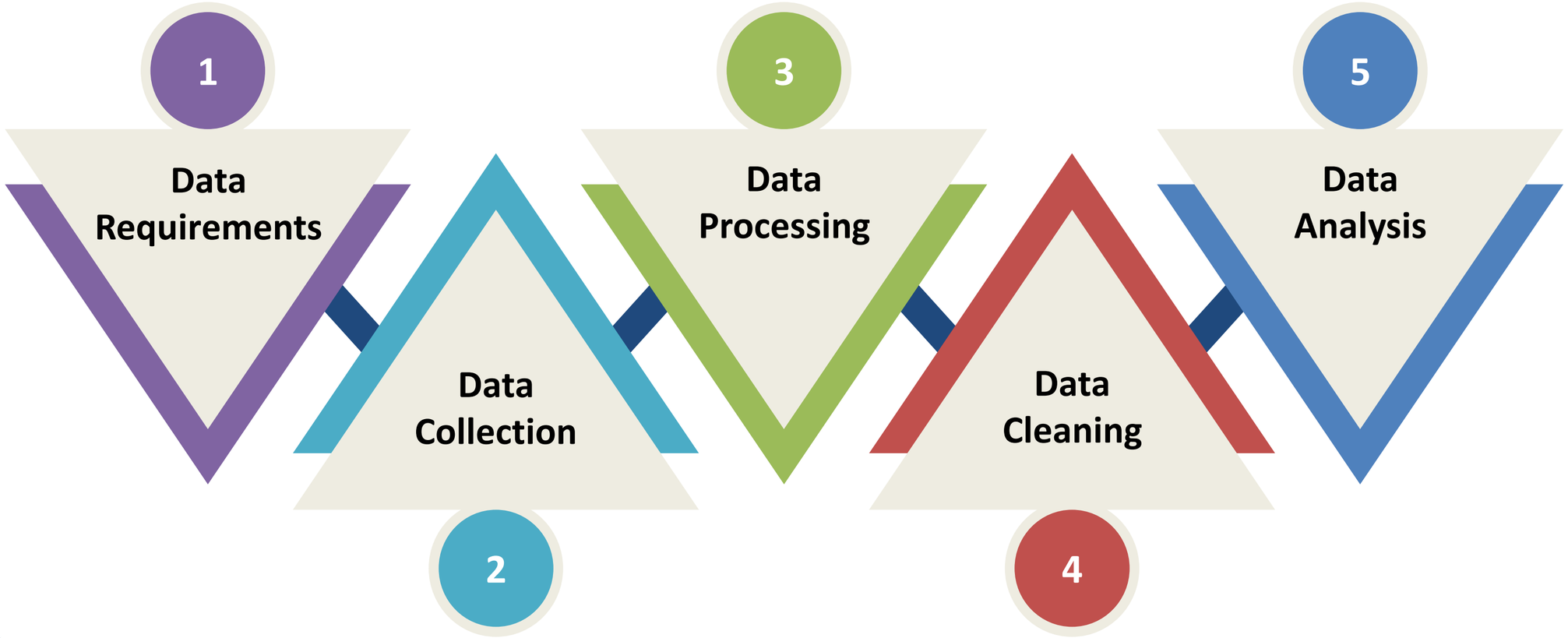
Diagnostic Analytics

Focuses more on why something happened

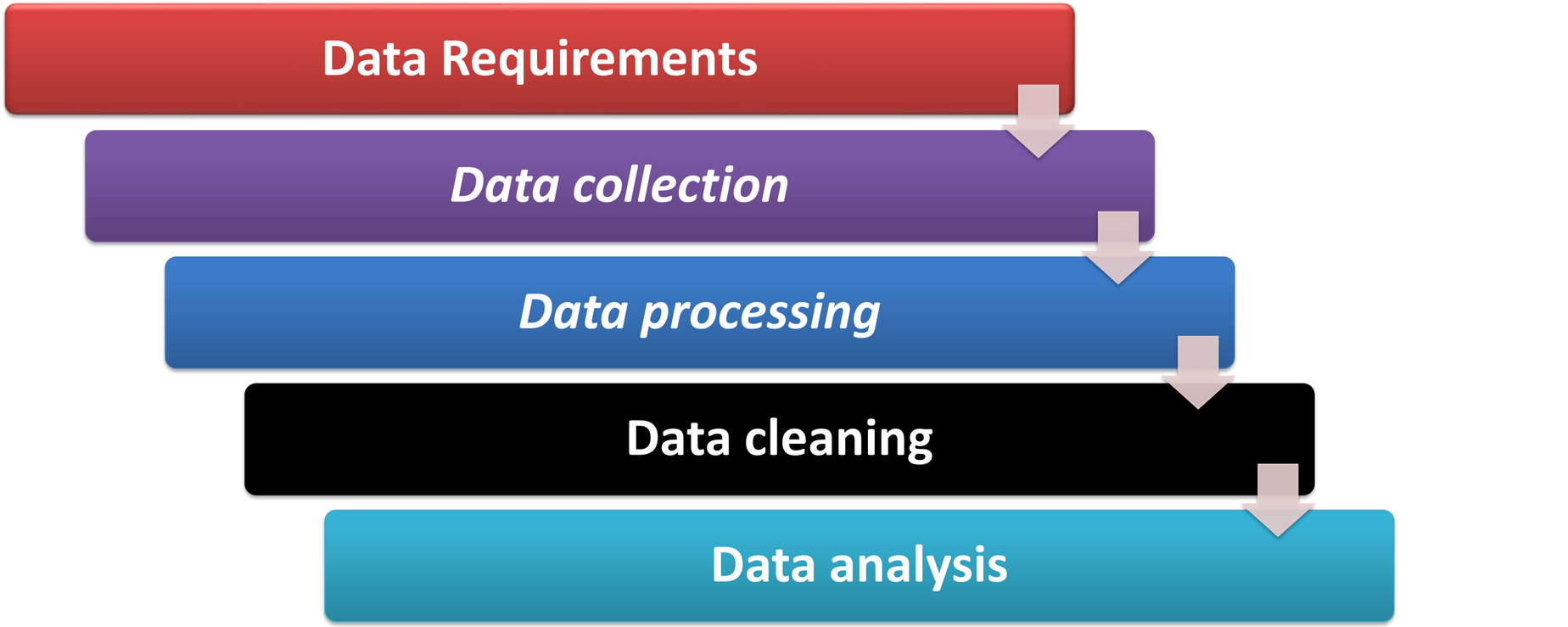
Prescriptive Analysis

Suggests a course of action

Data Analytics Process



Data Analytics Process



Data Analytics Process

Data Requirements	• Data are specified based upon the requirements • Specific variables regarding a population (age, income) • may be numerical or categorical
Data collection	• collected from a variety of sources (sensors, CCTV, satellite, recording devices) • It also be obtained through interviews and downloads from online sources
Data processing	• placing data into rows and columns in a table format for further analysis (spreadsheet /statistical software)
Data cleaning	• data may be incomplete, contain duplicates, or contain errors • is the process of preventing and correcting these errors • Tasks like record matching, deduplication, and column segmentation
Data Analysis	• variety of techniques referred • Mathematical formulas / models called algorithms may be applied to the data to identify relationships among the variables, such as correlation or causation

Data Analytics Comprises …

Business Analytics	• monitors the status of any relevant business component or characteristic on-demand, in real-time
Data Management	• It handles large amount of data, different data types including unstructured data
Predictive Analytics and Performance Management	• Helps to identify trends and characteristics, both positive and negative
Data warehousing	• It can handle the traditional, processed data, unprocessed, raw data along with live data streams
Business intelligence	• Analyze data and make it as useful business decision

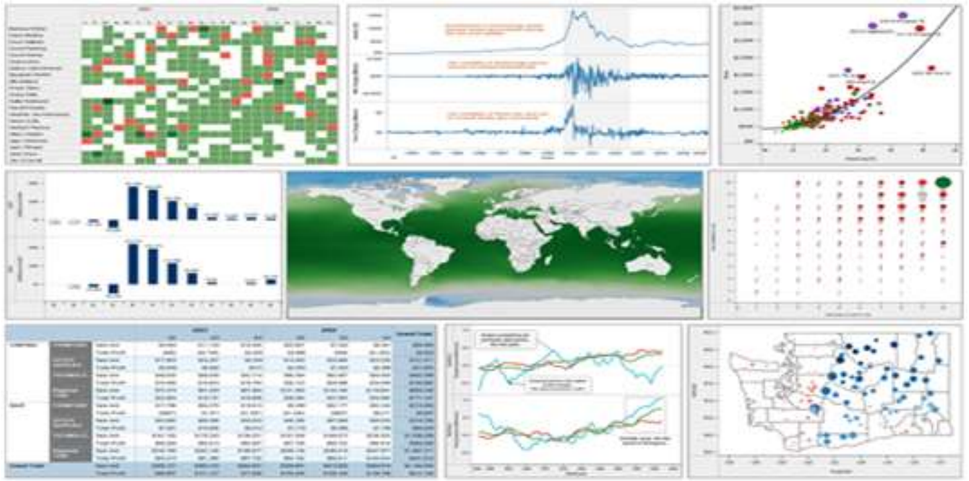
Modern Data Analytics Tools

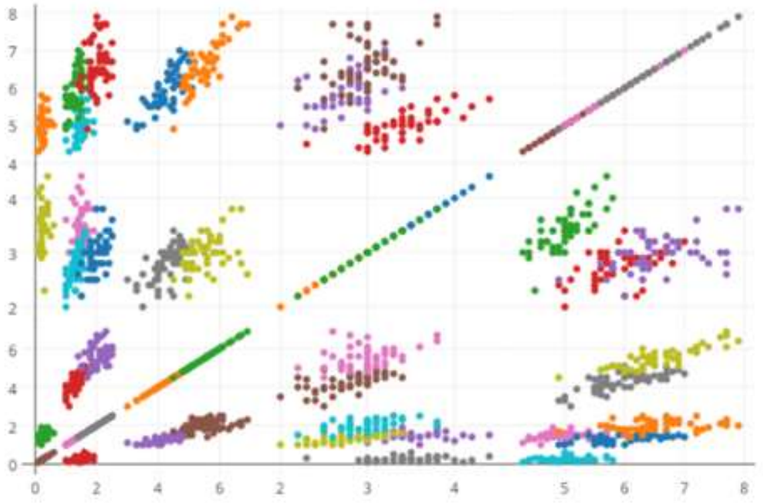
RapidMiner operates through visual programming and is capable of manipulating, analyzing and modeling data

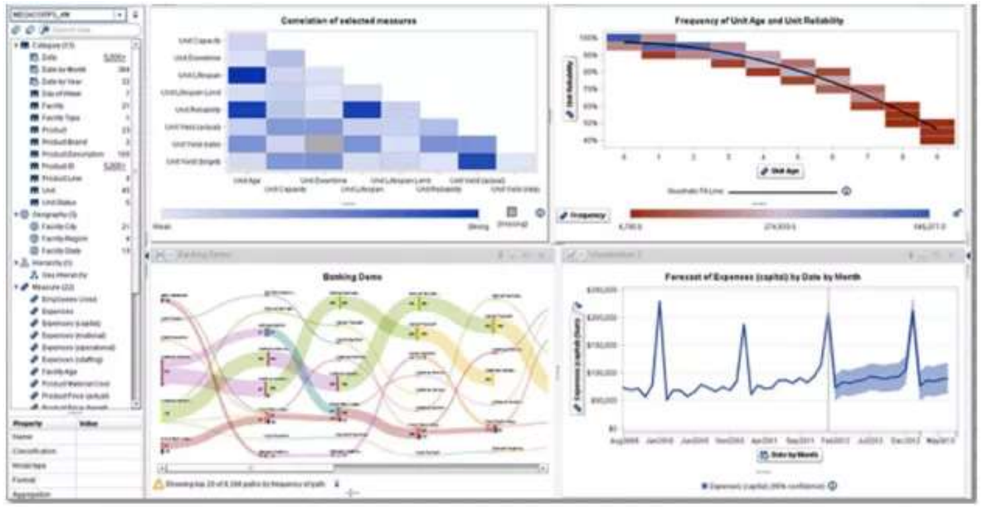
OpenRefine is a data cleaning software that allows you to get everything ready for analysis.

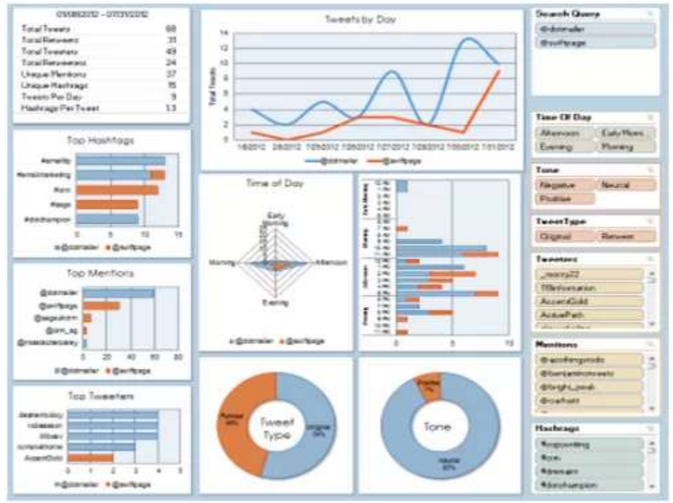
SAP Analytics

Modern Data Analytics Tools





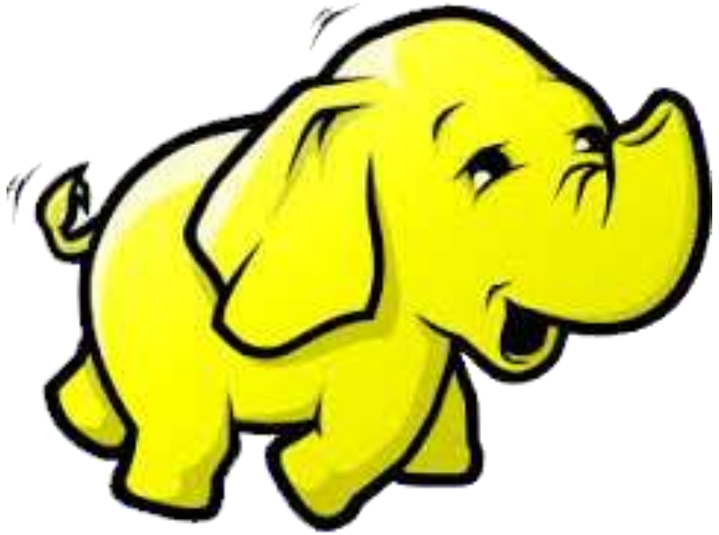




Modern Data Analytics Tool - Hadoop

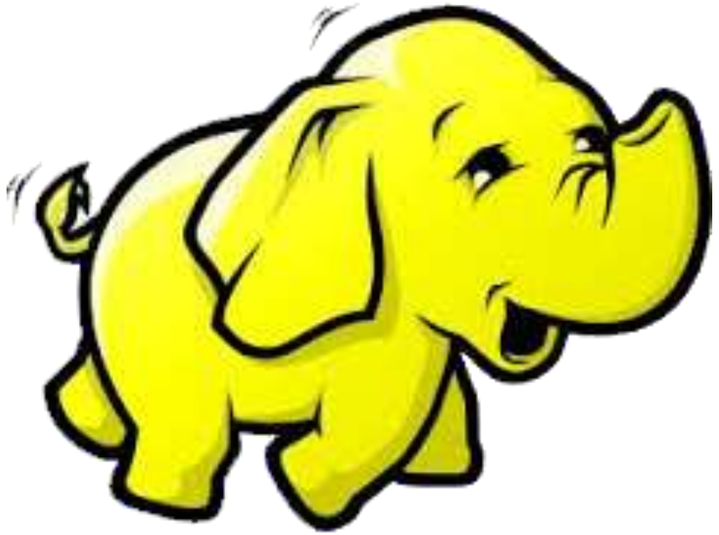
- Hadoop – Java based frame work allow for distributed processing of large data set using commodity hardware

- Open source data management with scale-out storage and distributed processing

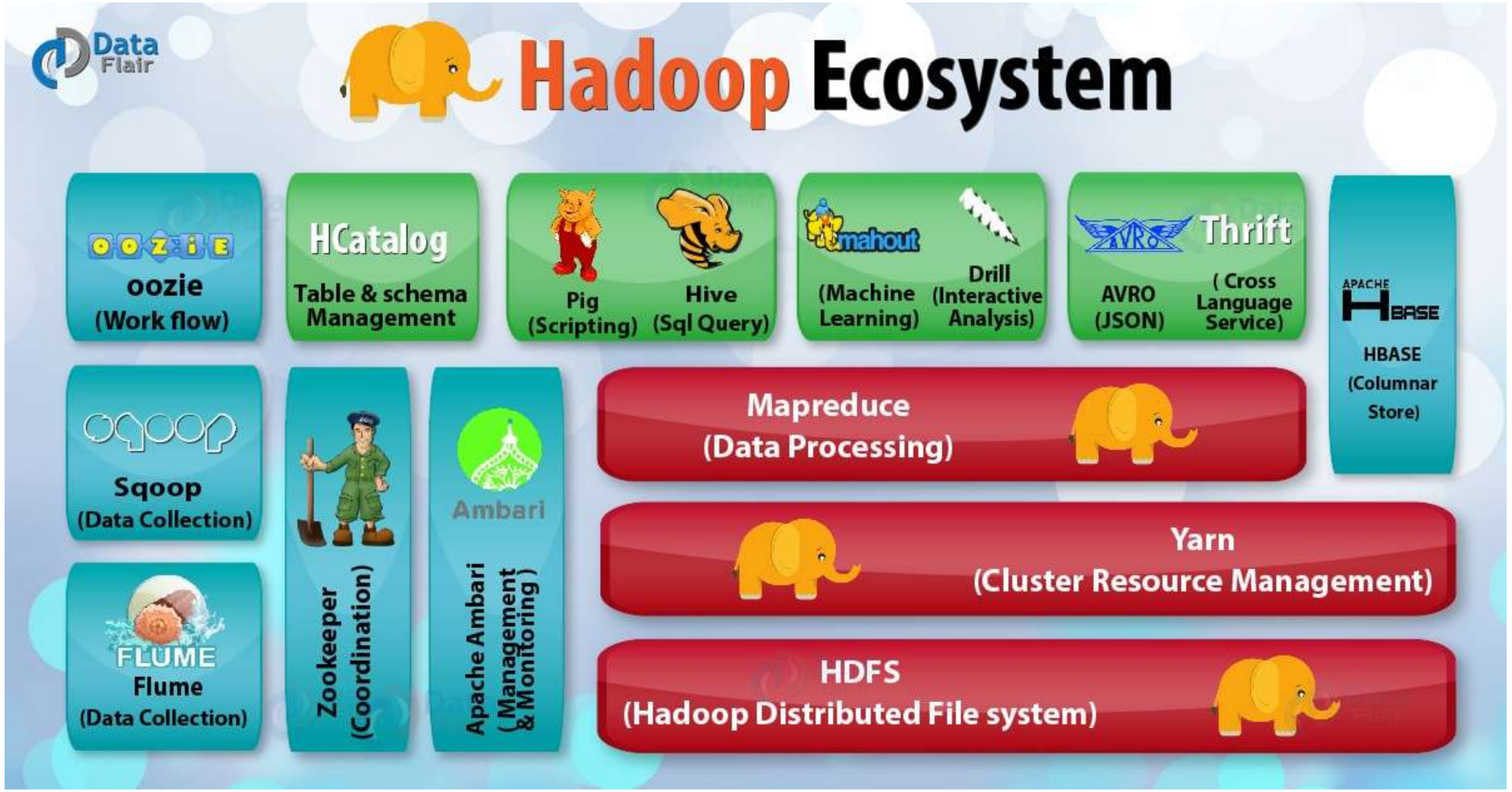


Modern Data Analytics Tool - Hadoop

“Hadoop is a technology to store massive datasets on a cluster of cheap machines in a distributed manner”



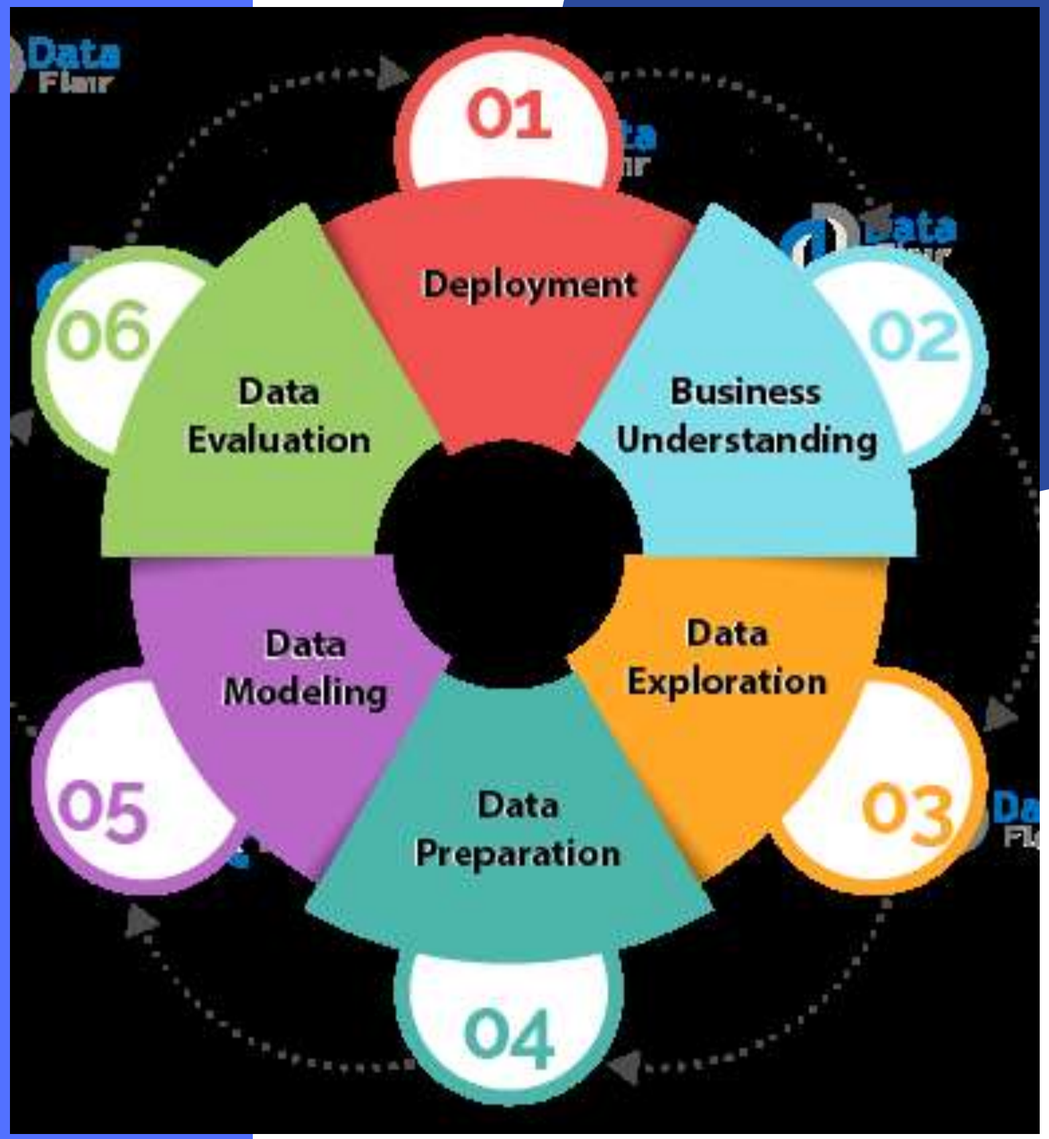
Modern Data Analytics Tool - Hadoop



ANALYTIC PROCESS TOOLS

There are 6 analytic processes:

1. Deployment
2. Business understanding
3. Data exploration
4. Data preparation
5. Data modeling
6. Data evaluation



STEP 1: DEPLOYMENT

- Here we need to: –
- plan the deployment and monitoring and maintenance, – we need to produce a final report and review the project.
- – In this phase,
    - we deploy the results of the analysis.
    - This is also known as reviewing the project.

STEP 2: BUSINESS UNDERSTANDING

- The very first step consists of business understanding.
- Whenever any requirement occurs, firstly we need to determine the business objective,
- assess the situation,
- determine data mining goals and then
- produce the project plan as per the requirement.

• Business objectives are defined in this phase.

STEP 3: DATA EXPLORATION

- The second step consists of Data understanding.
  - For the further process, we need to gather initial data, describe and explore the data and verify data quality to ensure it contains the data we require.
  - Data collected from the various sources is described in terms of its application and the need for the project in this phase.
  - This is also known as data exploration.
- This is necessary to verify the quality of data collected.

STEP4: DATA PREPARATION

From the data collected in the last step,

– we need to select data as per the need, clean it, construct it to get useful information and – then integrate it all.

- Finally, we need to format the data to get the appropriate data.
- Data is selected, cleaned, and integrated into the format finalized for the analysis in this phase.

STEP5: DATA MODELING

we need to – select a modeling technique, generate test design, build a model and assess the model built.

- The data model is build to

– analyze relationships between various selected objects in the data,

– test cases are built for assessing the model and model is tested and implemented on the data in this phase.

- Where processing is hosted?

– Distributed Servers / Cloud (e.g. Amazon EC2)

- Where data is stored?

– Distributed Storage (e.g. Amazon S3)

- What is the programming model?

– Distributed Processing (e.g. MapReduce)

STEP5: DATA MODELING

How data is stored & indexed?

– High-performance schema-free databases (e.g. MongoDB)

- What operations are performed on data?

– Analytic / Semantic Processing

- Big data tools for HPC and supercomputing

– MPI

- Big data tools on clouds

– MapReduce model

– Iterative MapReduce model

– DAG model

– Graph model

– Collective model

- Other BDA tools

– SaS

– R

– Hadoop

Thus the BDA tools are used through out the BDA applications development.