



# Introduction to Data Science

**RJSPIT101**



# Unit I

**Data Science Technology Stack:** Rapid Information Factory Ecosystem, Data Science Storage Tools, Data Lake, Data Vault, Data Warehouse Bus Matrix, Data Science Processing Tools, Spark, Mesos, Akka, Cassandra, Kafka, Elastic Search, R, Scala, Python.

**Layered Framework:** Definition of Data Science Framework, Cross Industry Standard Process for Data Mining (CRISP-DM), Homogeneous Ontology for Recursive Uniform Schema, The Top Layers of a Layered Framework, Layered Framework for High-Level Data Science and Engineering.

**Business Layer:** Business Layer, Engineering a Practical Business Layer

## Unit II

**Utility Layer:** Basic Utility Design, Engineering a Practical Utility Layer.

**Three Management Layers:** Operational Management Layer, Processing-Stream Definition and Management, Audit, Balance, and Control Layer, Balance, Control, Yoke Solution, Cause-and-Effect, Analysis System, Functional Layer, Data Science Process.

**Retrieve Superstep:** Data Lakes, Data Swamps, Training the Trainer Model, Understanding the Business Dynamics of the Data Lake, Actionable Business Knowledge from Data Lakes, Engineering a Practical Retrieve Superstep, Connecting to Other Data Sources.

# Unit III

**Assess Superstep:** Assess Superstep, Errors, Analysis of Data, Practical Actions, Engineering a Practical Assess Superstep.

**Process Superstep:** Data Vault, Time-Person-Object-Location-Event Data Vault, Data Science Process, Data Science.

# Unit IV

**Transform Superstep:** Transform Superstep, Building a Data Warehouse, Transforming with Data Science, Hypothesis Testing, Overfitting and Underfitting, Precision-Recall, Cross-Validation Test, Univariate Analysis, Bivariate Analysis, Multivariate Analysis, Linear Regression, Logistic Regression, Clustering Techniques, ANOVA, Principal Component Analysis (PCA), Decision Trees, Support Vector Machines, Random Forests.

# Practical List

- NumPy, Pandas, Matplotlib and Seaborn Basics.
- Collecting and loading structured and unstructured data.
- Using Data Wrangling processes: Data discovery, data pre-processing, data validation etc. for various types of data.
- Basic utility design, Data auditing and Exploratory Data Analysis.
- Retrieve Superstep.
- Access Superstep.
- Processing Data.
- Transforming Data:  
Using Machine Learning Algorithms.
- Organizing and Generating data.
- Data Visualization.

## References

- Andreas François Vermeulen, “Practical Data Science A Guide to Building the Technology Stack for Turning Data Lakes into Business Assets”, Apress, 2018.
- Sinan Ozdemir, “Principles of Data Science”, PACKT, 2016
- Peter Bruce, Andrew Bruce, “Practical Statistics for Data Science”, O’Reilly, 2017.
- Wes McKinney, “Python for Data Analysis: Data Wrangling with Pandas, NumPy and IPython”, O’Reilly, 2nd Edition.
- Allen B. Downey, “Think Stats : Probability and Statistics for Programmers”, Green Tea Press.
- Jose Unpingco, “Python for Probability, Statistics and Machine Learning”, Springer.



# Introduction to Data Science



# Data All Around

- ◆ Lots of data is being collected and warehoused
  - Web data, e-commerce
  - Financial transactions, bank/credit transactions
  - Online trading and purchasing
  - Social Network



# Types of Data

- ◆ Relational Data  
(Tables/Transaction/Legacy Data)
- ◆ Text Data (Web)
- ◆ Semi-structured Data (XML)
- ◆ Graph Data
- ◆ Social Network, Semantic Web (RDF), ...
- ◆ Streaming Data
- ◆ You can afford to scan the data once

# What to with these Data

- ◆ Aggregation and Statistics
  - Data warehousing and OLAP
- ◆ Indexing, Searching, and Querying
  - Keyword based search
  - Pattern matching (XML/RDF)
- ◆ Knowledge discovery
  - Data Mining
  - Statistical Modeling

# What is Data Science?

An area that manages, manipulates, extracts, and interprets knowledge from tremendous amount of data

Data science (DS) is a multidisciplinary field of study with goal to address the challenges in big data

Data science principles apply to all data – big and small

Theories and techniques from many fields and disciplines are used to investigate and analyze a large amount of data to help decision makers in many industries such as science, engineering, economics, politics, finance, and education

## Computer Science

Pattern recognition, visualization, data warehousing, High performance computing, Databases, AI

## Mathematics

Mathematical Modeling

## Statistics

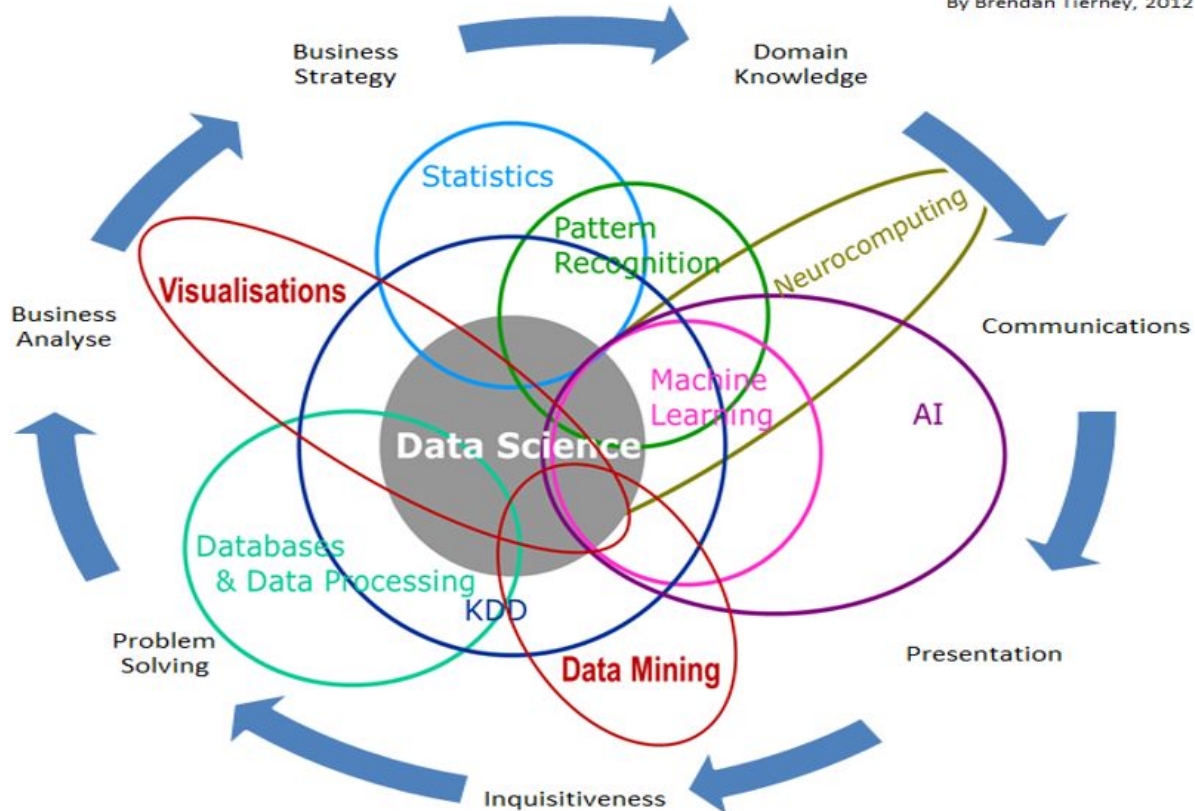
Statistical and Stochastic modeling, Probability.

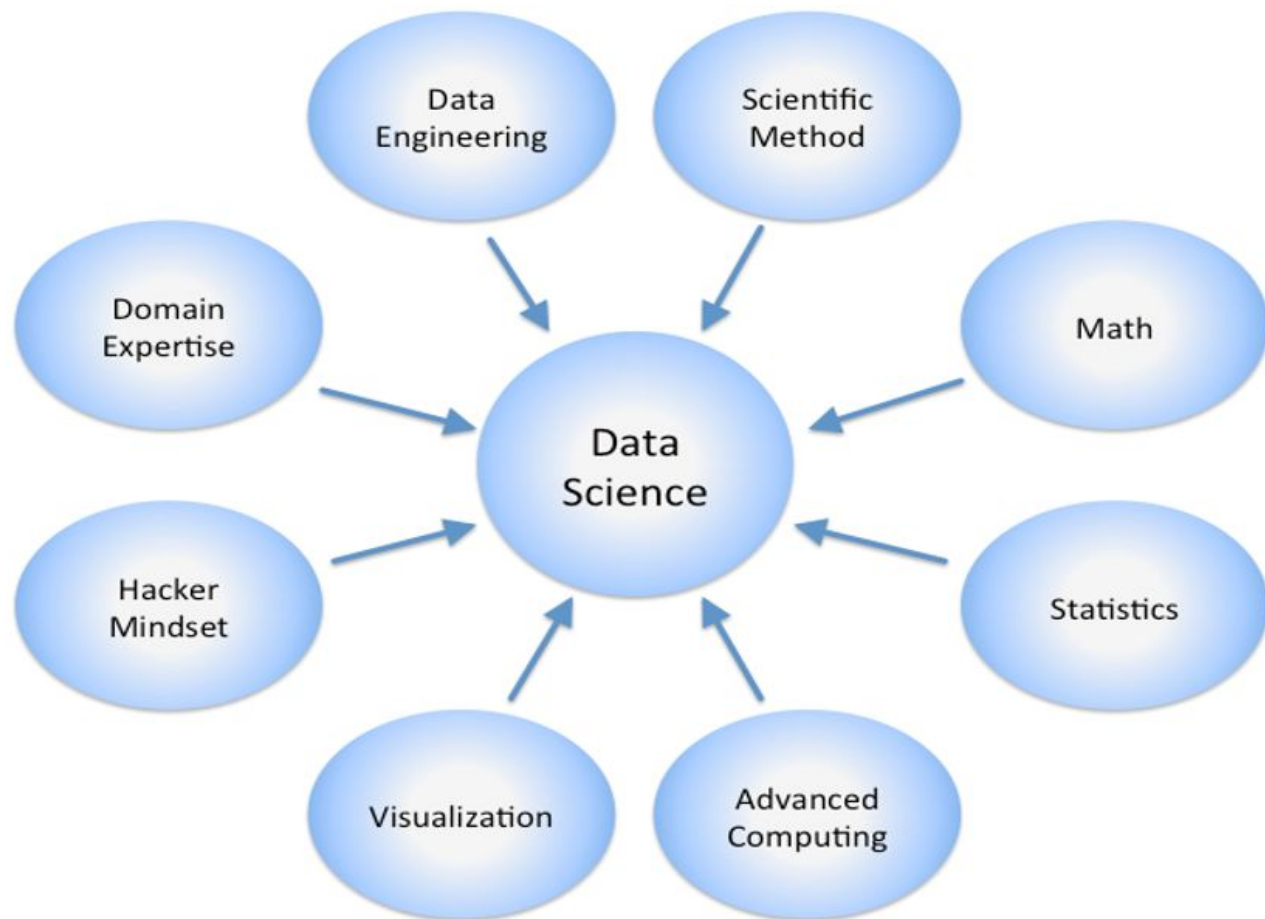
# Data Science

- In 1960, Peter Naur started using the term data science as a substitute for computer science.
- He stated that to work with data, you require more than just computer science.
- Data science is an **interdisciplinary science** that incorporates practices and methods with actionable knowledge and insights from data in heterogeneous schemas (structured, semi-structured, or unstructured).
- It amalgamates the scientific fields of data exploration with thought-provoking research fields such as data engineering, information science, computer science, statistics, artificial intelligence, machine learning, data mining, and predictive analytics.

# Data Science Is Multidisciplinary

By Brendan Tierney, 2012







*...solving problems with data...*



*...sounds cool!*

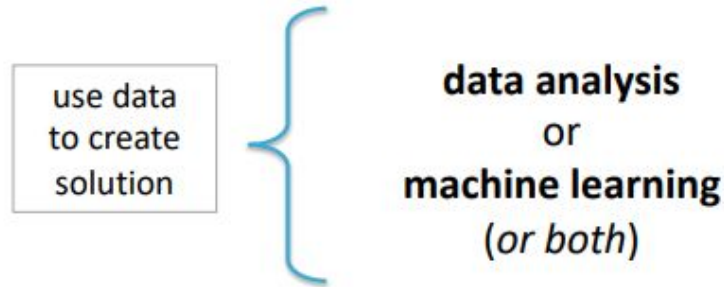
*What makes a good data scientist?*



## *...solving problems with data...*



## *...which step is most challenging?*



# WHAT IS DATA ANALYSIS?

*...using data to discover useful information...*



- **data:** anything you can *measure* or *record*



- **statistics:** summarize (and visualize) *main characteristics* of the data



- **algorithms:** apply algorithms to find *patterns* in the data

# WHAT IS MACHINE LEARNING?

*...creating and using models that learn from data...*



- **data:** anything you can *measure* or *record*



- **model:** specification of a (mathematical) *relationship* between different variables



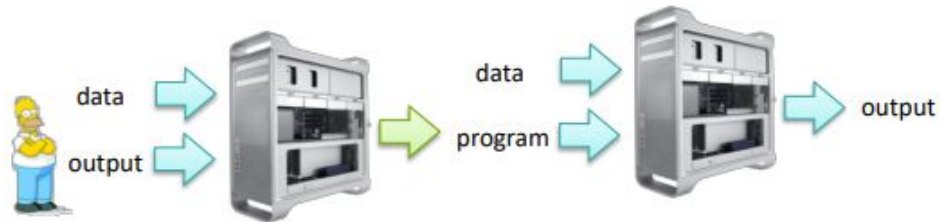
- **evaluation:** how well does the model *work?*
-

# WHAT IS MACHINE LEARNING?

- Traditional CS



- Machine Learning



# WHAT IS MACHINE LEARNING?

*...creating and using models that learn from data...*

## Examples

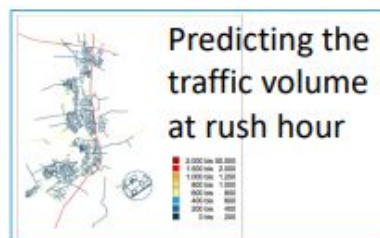
Identifying zip code  
from handwritten  
digits



Detecting  
communities  
in social  
networks



Predicting the  
traffic volume  
at rush hour



Detecting fraudulent  
credit card  
transactions



Determining the  
location of distribution  
centers based on  
customers' residence



# Data Science...

Let us discuss the vital elements from these sciences that you will use to process your data lake into actionable knowledge.

Also we will discuss the recognized science procedures for data lakes.

These core skills are a key set of assets to perfect as you begin your encounters with data science.

# Data Analytics

- Data analytics is the science of fact-finding analysis of raw data, with the goal of drawing conclusions from the data lake.
- Data analytics is driven by certified algorithms to statistically define associations between data that produce insights.
- The perception of certified algorithms is exceptionally significant when you want to convince other business people of the importance of the data insights you have gleaned.



# Machine Learning

- The business world is buzzing with activities and ideas about machine learning and its application to numerous business environments.
- Machine learning is the capability of systems to learn without explicit software development. It evolved from the study of pattern recognition and computational learning theory.
- The impact is that, with the appropriate processing and skills, you can augment your own data capabilities.
- Training enables a processing environment to complete several magnitudes of discoveries in the time it takes to have a cup of coffee.
- This skill is an essential part of achieving major gains in shortening the data-to-knowledge cycle

# Data Mining

- Data mining is processing data to isolate patterns and establish relationships between data entities within the data lake.
- For data mining to be successful, there is a small number of critical data-mining theories that you must know about data patterns.

# Statistics

- Statistics is the study of the collection, analysis, interpretation, presentation, and organization of data. Statistics deals with all aspects of data, including the planning of data collection, in terms of the design of surveys and experiments.
- Data science and statistics are closely related. I will show you how to run through series of statistics models covering data collection, population, and samples to enhance your data science deliveries.

# Algorithms

- An algorithm is a self-contained step-by-step set of processes to achieve a specific outcome. Algorithms execute calculations, data processing, or automated reasoning tasks with repeatable outcomes.
- Algorithms are the backbone of the data science process.
- You should assemble a series of methods and procedures that will ease the complexity and processing of your specific data lake.

# Data Visualization

- Data visualization is your key communication channel with the business. It consists of the creation and study of the visual representation of business insights. Data science's principal deliverable is visualization.
- You will have to take your highly technical results and transform them into a format that you can show to non-specialists.
- The successful transformation of data results to actionable knowledge is a skill set I will cover in detail in later chapters.
- If you master the visualization skill, you will be most successful in data science.

# Story Telling

- Data storytelling is the process of translating data analyses into layperson's terms, in order to influence a business decision or action.
- You can have the finest data science, but without the business story to translate your findings into business-relevant actions, you will not succeed.




Thank You...!!!



# Introduction: Basic Terms

**Data Science Technology Stack:** Rapid Information Factory Ecosystem, Data Science Storage Tools, Data Science Warehouse Bus Matrix, Data Science Processing Tools, Spark, Mesos, Akka, Cassandra, Kafka, Elastic Search





# Introduction

The Data Science Technology Stack covers the data processing requirements in the Rapid Information Factory ecosystem.

## Topics

- Recognize the basics of data science tools and their influence on modern data lake development.
- Discover the techniques for transforming a data vault into a data warehouse bus matrix.
- Use of Spark, Mesos, Akka, Cassandra, and Kafka, to tame your data science requirements.
- Use of elastic search and MQTT (MQ Telemetry Transport), to enhance your data science solutions.
- Recognize the influence of R as a creative visualization solution.
- Introduce the impact and influence on the data science ecosystem of such programming languages as R, Python, and Scala.

# Rapid Information Factory Ecosystem

- The Rapid Information Factory ecosystem is a convention of techniques I use for my individual processing developments.
- The tools can be used in any configuration or permutation that is suitable to your specific ecosystem.
- I recommend that you begin to formulate an ecosystem of your own or simply adopt mine.
- As a prerequisite, you must become accustomed to a set of tools you know well and can deploy proficiently.

# Data Science Storage Tools

- This data science ecosystem has a series of tools that you use to build your solutions.
- This environment is undergoing a rapid advancement in capabilities, and new developments are occurring every day
- We will discuss the tools that are used to perform practical data science and the basic data methodologies.

# Schema-on-Write and Schema-on-Read

There are two basic methodologies that are supported by the data processing tools.

1. Schema-on-Write Ecosystem
2. Schema-on-Read Ecosystem

# Schema-on-Write Ecosystem

- A traditional relational database management system (RDBMS) requires a schema before you can load the data.
- To retrieve data from my structured data schemas, you may have been running standard SQL queries for a number of years.
- **Benefits** include the following:
  - In traditional data ecosystems, tools assume schemas and can only work once the schema is described, so there is only one view on the data.
  - The approach is extremely valuable in articulating relationships between data points, so there are already relationships configured.
  - It is an efficient way to store “dense” data.
  - All the data is in the same data store

# Schema-on-Write Ecosystem...

- Schema-on-write isn't the answer to every data science problem.
- Among the **downsides** of this approach are that
  - Its schemas are typically purpose-built, which makes them hard to change and maintain.
  - It generally loses the raw/atomic data as a source for future analysis.
  - It requires considerable modeling/implementation effort before being able to work with the data.
  - If a specific type of data can't be stored in the schema, you can't effectively process it from the schema.

# Schema-on-Read Ecosystem

- This alternative data storage methodology does not require a schema before you can load the data. Fundamentally, you store the data with minimum structure. The essential schema is applied during the query phase.
- **Benefits** include the following:
  - It provides flexibility to store unstructured, semi-structured, and disorganized data.
  - It allows for unlimited flexibility when querying data from the structure.
  - Leaf-level data is kept intact and untransformed for reference and use for the future.
  - The methodology encourages experimentation and exploration.
  - It increases the speed of generating fresh actionable knowledge.
  - It reduces the cycle time between data generation to availability of actionable knowledge.

I recommend a hybrid between schema-on-read and schema-on-write ecosystems for effective data science and engineering

# Data Lake



A Data Lake is storage repository of large amount of raw data that means structure, semi-structure, unstructured data.

- This is the place where you can store three types of data structure, semi-structure, unstructured data with no fix amount of limit and storage to store the data.
- If we compare schema on write and data lake then we will find that schema on write store the data into the data warehouse in predefined database on the other hand data lake store the less data structure to store the data into the database.
- Data Lake follow to store less data into the structure database because it follows the schema on read process architecture to store the data.
- Data Lake allow us to transform the raw data that means structure, semi-structure, unstructured data into the structure data format so that SQL query could be performed for the analysis.
- Most of the time data lake is deployed by using the distributed data object storage database which enable the schema on read so that business analytics and data mining tools and algorithms can be applied on the data.
- Retrieval of data is so fast because there is no schema applied. Data must be access without any failure or any complex reason.
- Data Lake is similar to real time river or lake where the water comes from different- different places and at the last all the small- small river and lake are merged into the big river or lake where large amount of water are stored, whenever there is need of water then it can be used by anyone.
- It is low cost and effective way to store the large amount of data stored into centralized database for further organizational analysis and deployment.



# Data Vault

- Data Vault is a database modeling method which is designed to store the long-term historical storage amount of data and it can be controlled by using the data vault.
- In Data Vault, data must come from different sources and it is designed in such a way that data could be loaded in parallel ways so that large amount of data implementation can be done without any failure or any major design.
- Data Vault is the process of transforming the schema on read data lake into schema on write data lake.
- Data Vault are designed schema on read query request and after that it would be converted into the data lake because schema on read increases the speed of generating new data for the better analysis and implementation.
- Data Vault store a single version of data and does not distinguish between good data and bad data.
- Data Lake and Data Vault are built by using the three main components or structure of data i.e. Hub, Link and satellite.

# Hub

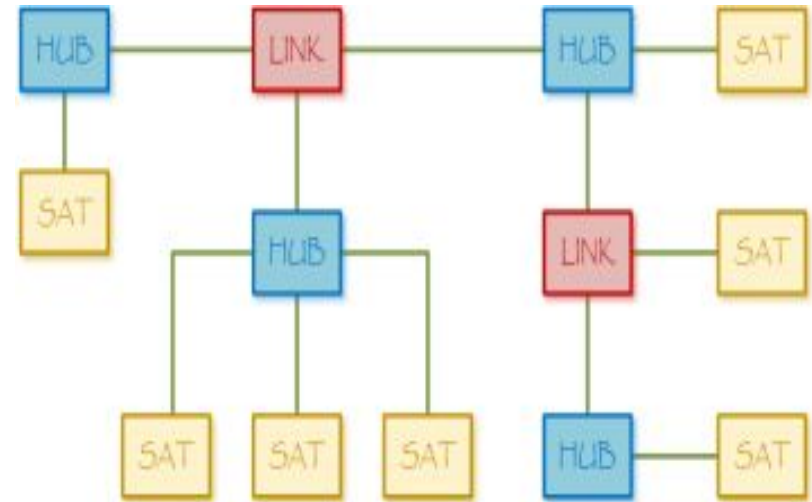
- Hub has unique business key with low amount of data to be changed and meta data that means data is the main source of generating the hubs.
- Hub contains surrogate key for each metadata information and each hub items i.e. origin of this business key.
- Hub contains a set of unique business key that will never change over a period manner.
- There are different types of hubs like person hub, time hub, object hub, event hub, locations hub. The Time hub contains ID Number, ID Time Number, ZoneBasekey, DateTime key, DateTimeValue and all these links are interconnected to each other like Time-Person, Time-Object, Time-Event, Time-Location, Time-Links etc.
- The Person hub contains IDPerson, Number, FirstName, SecondName, LastName, Gender, TimeZone, BirthDateKey, BirthDate and all these links are interconnected to each other like Person-Time, Person Object, Person-Location, Person-Event, Person-Link etc.
- The Object hub contains IDObjectNumber, ObjectBaseKey, ObjectNumber, ObjectValue and all these links are interconnected to each other like Object-Time, Object-Link, Object-Event, Object Location, Object-Person etc.
- The Event hub contains ID Event Number, Event Type, Event Description and all these links are interconnected to each other like Event-Person, Event-Location, Event-Object, Event-Time etc.
- The Location hub contains ID Location Number, Object Base Key, Location Number, Location Name, Longitude and Latitude all these links are interconnected to each other like Location-Person, Location Time, Location-Object, Location-event etc.

# Link

- Link plays a very important role during transaction and association of business key. The Table relate to each other depending upon the attribute of table like that one to one relationship, one to many relationships, Many to One relationship, Many to many relationships.
- Link represent and connect only element in the business relationships because when one node or link relate to one or another link on that time data transfers smoothly.

# Satellite

- When the hubs and links produce and form the structure of satellites which store no chronological structure of data means then it would not provide the information about the mean, median, mode, maximum, minimum, sum of the data.
- Satellites are the strong structure of data that store a detailed information about the related data or business characteristics key and stores large volume of data vault.
- The combinations of all these three i.e. hub, link, and satellites are formed together to help the data analytics and data scientists and data engineer to store the business structure, types of information or data into it.



# Data Warehouse Bus Matrix

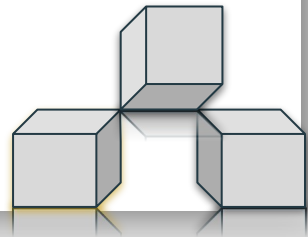
The Enterprise Bus Matrix is a data warehouse planning tool and model created by Ralph Kimball and used by numerous people worldwide over the last 40+ years.



The bus matrix and architecture builds upon the concept of conformed dimensions that are interlinked by facts.

The data warehouse is a major component of the solution required to transform data into actionable knowledge.

# Data Science Processing Tools

Data Science Tools transform your data lakes into data vaults and then into data warehouses. These tools are the workhorses of the data science and engineering ecosystem.





## Chapter 2: Vermeulen-Krennwallner- Hillman-Clark

# Vermeulen-Krennwallner-Hillman-Clark

- Vermeulen-Krennwallner-Hillman-Clark Group (VKHCG) is a hypothetical medium-size international company. It consists of four subcompanies:
  - Vermeulen PLC
  - Krennwallner AG
  - Hillman Ltd
  - Clark Ltd.
-



# Vermeulen PLC