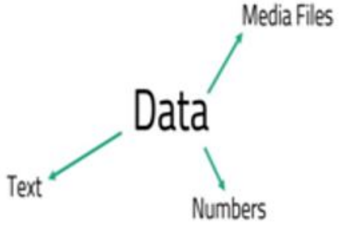
BIG DATA

Introduction to Big Data

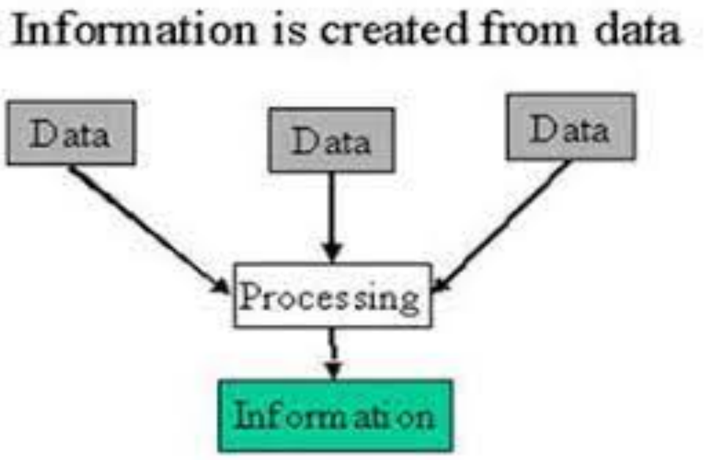
- – Big Data Overview
- – Background of Data Analytics
- – Role of Distributed System in Big Data
- – Role of Data Scientist
- – Current Trend in Big Data Analytics

Data and Information

- **Data** is defined as facts or figures



- **Information** is a processed, organized data which gives logical meaning



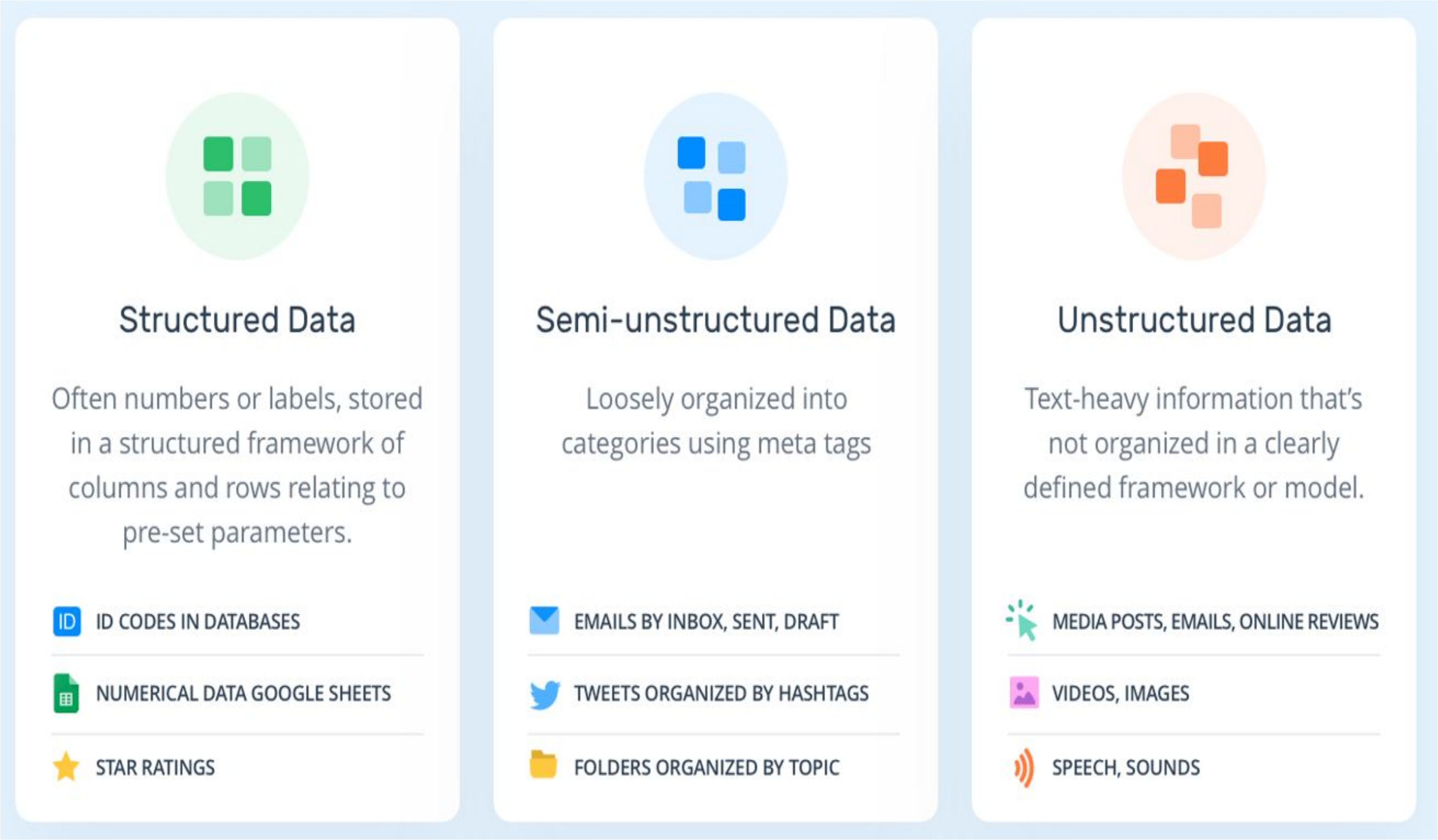
What is Big Data?

- **Big Data** is a collection of data that is huge in volume, yet growing exponentially with time. It is a data with so large size and complexity that none of traditional data management tools can store it or process it efficiently. Big data is also a data but with huge size.
- Big data is a combination of **structured, semi structured and unstructured** data collected by organizations that can be mined for information and used in machine learning projects, predictive modeling and other advanced analytics applications.

Example of Big Data

- **Social networking sites:** Facebook, Twitter, LinkedIn all these sites generates huge amount of data on a day to day basis as they have billions of users worldwide.
- **E-commerce site:** Sites like Amazon, Flipkart, Alibaba generates huge amount of logs from which users buying trends can be traced.
- **Weather Station:** All the weather station and satellite gives very huge data which are stored and manipulated to forecast weather.
- **Telecom company:** Telecom giants like Airtel, Vodafone study the user trends and accordingly publish their plans and for this they store the data of its million users.
- **Share Market:** Stock exchange across the world generates huge amount of data through its daily transaction.
- A single **Jet engine** can generate ***10+terabytes*** of data in ***30 minutes*** of flight time. With many thousand flights per day, generation of data reaches up to many ***Petabytes.***

Types of Big Data



Structured Data

Often numbers or labels, stored in a structured framework of columns and rows relating to pre-set parameters.

- ID CODES IN DATABASES
- NUMERICAL DATA GOOGLE SHEETS
- STAR RATINGS

Semi-unstructured Data

Loosely organized into categories using meta tags

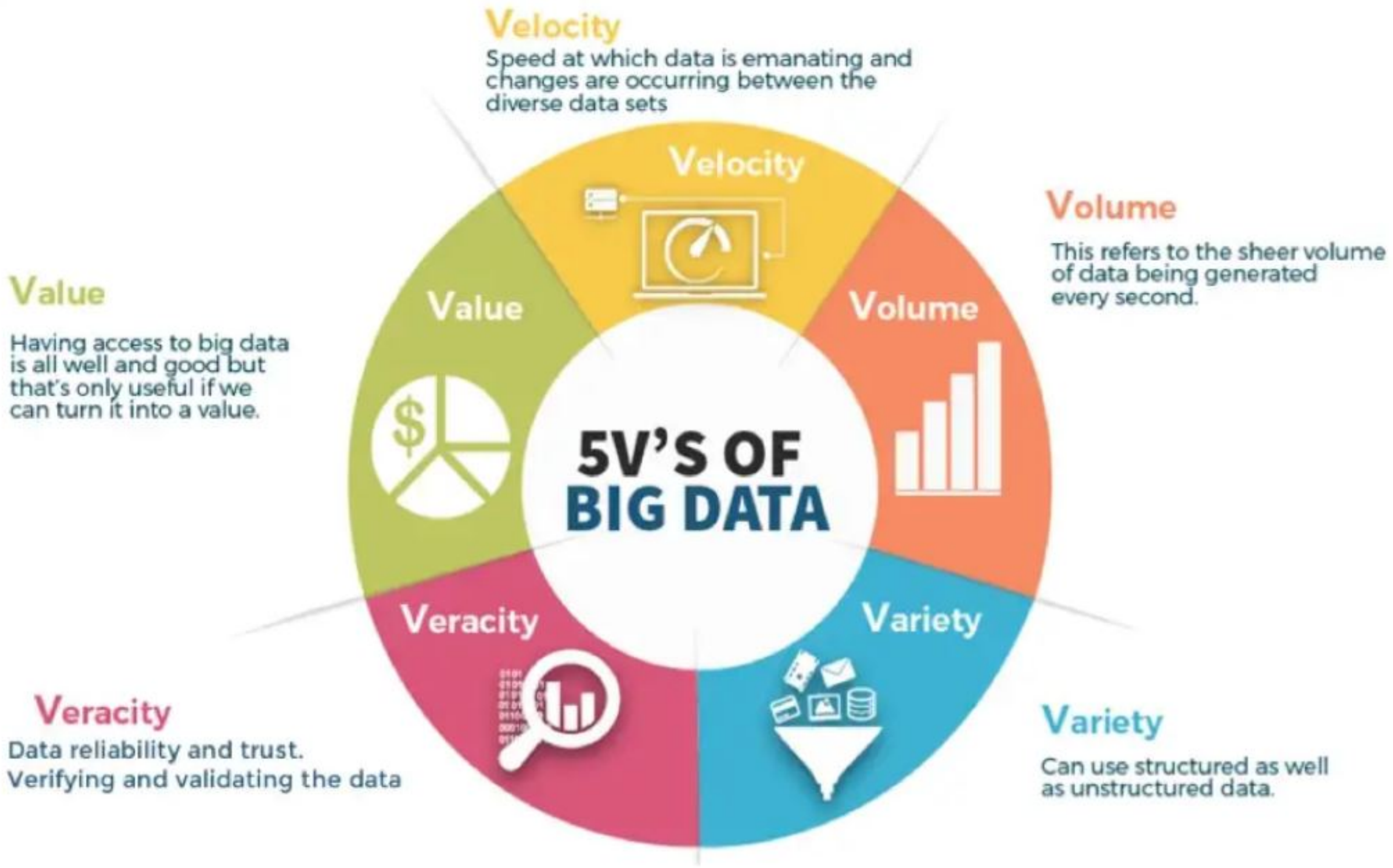
- EMAILS BY INBOX, SENT, DRAFT
- TWEETS ORGANIZED BY HASHTAGS
- FOLDERS ORGANIZED BY TOPIC

Unstructured Data

Text-heavy information that's not organized in a clearly defined framework or model.

- MEDIA POSTS, EMAILS, ONLINE REVIEWS
- VIDEOS, IMAGES
- SPEECH, SOUNDS

Characteristics of Big Data



Velocity

Speed at which data is emanating and changes are occurring between the diverse data sets

Volume

This refers to the sheer volume of data being generated every second.

Value

Having access to big data is all well and good but that's only useful if we can turn it into a value.

Veracity

Data reliability and trust.
Verifying and validating the data

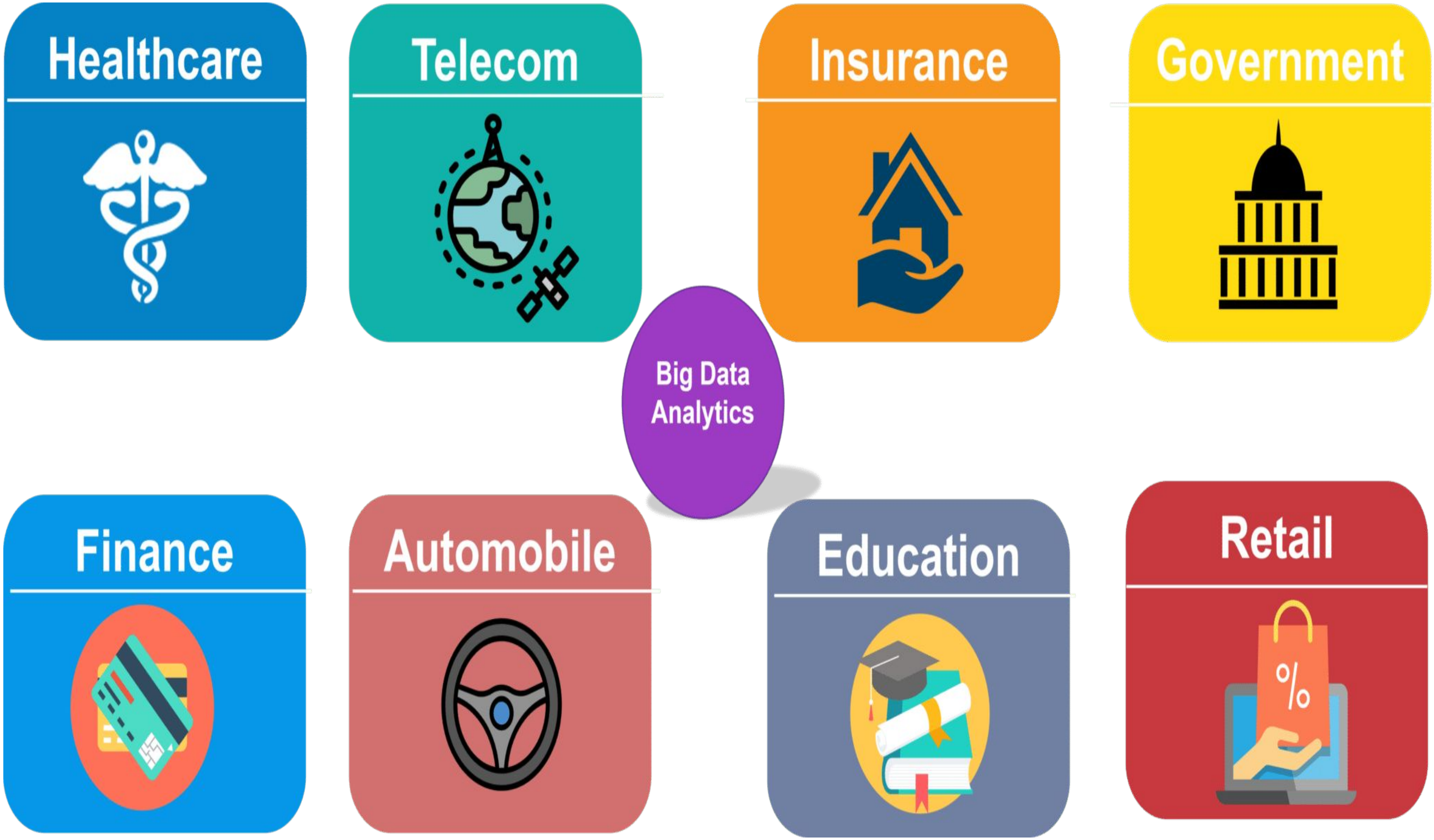
Variety

Can use structured as well as unstructured data.

- **Volume:** refers to the unimaginable amounts of information generated every second from social media, cell phones, cars, credit cards, images, video, and whatnot. **Facebook alone can generate about billion messages, 4.5 billion times that the “like” button is recorded, and over 350 million new posts are uploaded each day.**
- **Variety:** Data comes in all types of formats — from **structured**, numeric data in traditional databases to **unstructured** text documents, emails, videos, audios, stock ticker data, and financial transactions.

- **Veracity:-** refers to the quality of the data. Because data comes from so many different sources, it’s difficult to link, match, cleanse, and transform data across systems. Such an example of this is data generated from Medical experiments or trials.
- **Value:** is actually the amount of **valuable, reliable, and trustworthy** data that needs to be stored, processed, analyzed to find insights.
- **Velocity:** is defined as at which speed the data is generated from the source. An example of Velocity in Big Data is Tweets on Twitter or posts on Facebook.

Applications of Big Data



Healthcare

Telecom

Insurance

Government

Big Data Analytics

Finance

Automobile

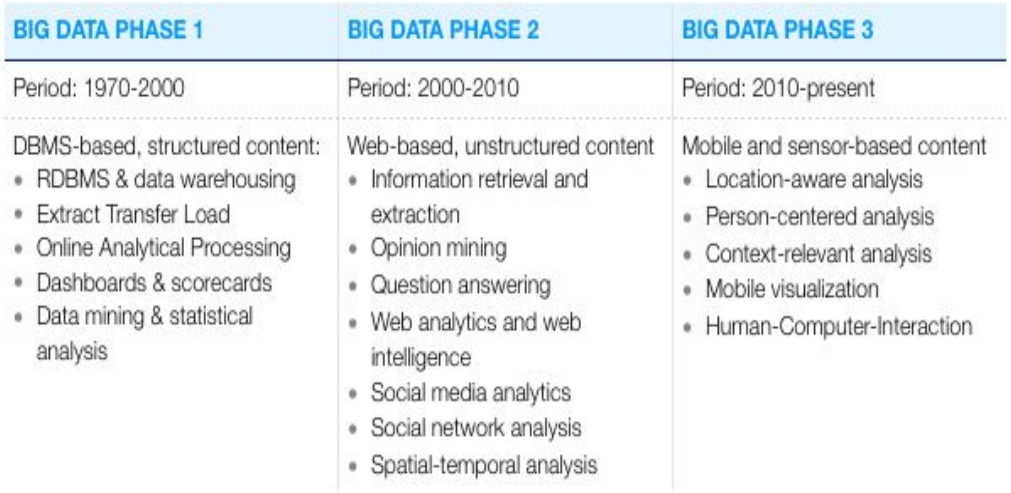
Education

Retail

Applications of Big Data

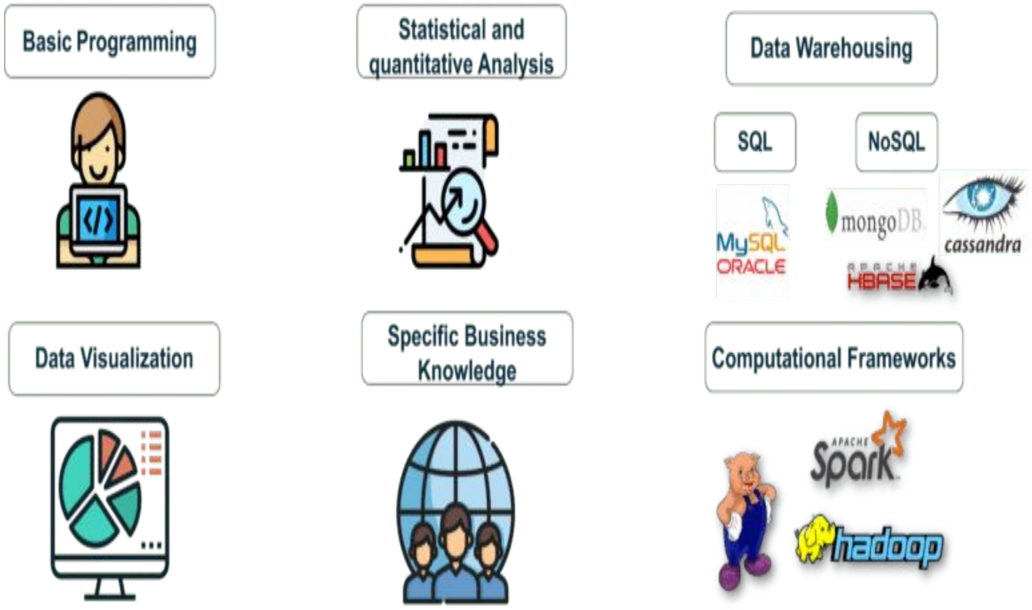
- Organizations from the different domains are investing in Big Data applications, for examining large data sets to uncover all hidden patterns, unknown correlations, market trends, customer preferences, and other useful business information.

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 Tools



Basic Programming

Statistical and quantitative Analysis

Data Warehousing

SQL

NoSQL

Data Visualization

Specific Business Knowledge

Computational Frameworks

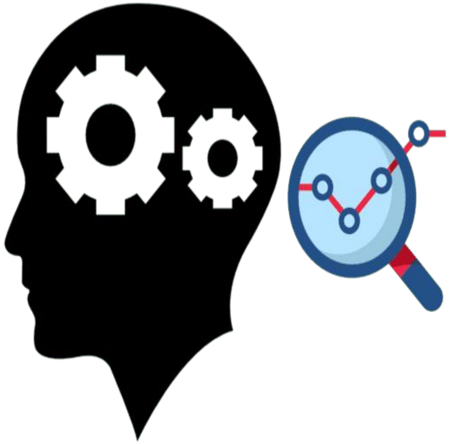
Big Data Analytics Tools

- **Hadoop** - helps in storing and analyzing data
- **MongoDB** - used on datasets that change frequently
- **Talend** - used for data integration and management
- **Cassandra** - a distributed database used to handle chunks of data
- **Spark** - used for real-time processing and analyzing large amounts of data
- **STORM** - an open-source real-time computational system
- **Kafka** - a distributed streaming platform that is used for fault-tolerant storage
- **Databricks** - combines data warehouses & data lakes into a lakehouse architecture. Collaborate on all data analytics & AI workloads

Big Data Use Case Example 1:

Starbucks uses behavioural analytics to cater to its customers

Starbucks gather a lot of info about their customers' coffee-buying habits from their preferred drinks to what time of day they're usually ordering



The company directs exciting offers and coupons to their customers and ensures to maintain their interest

Big Data Use Case Example 2:

P&G uses Market Basket Analysis and price optimization to optimize their products

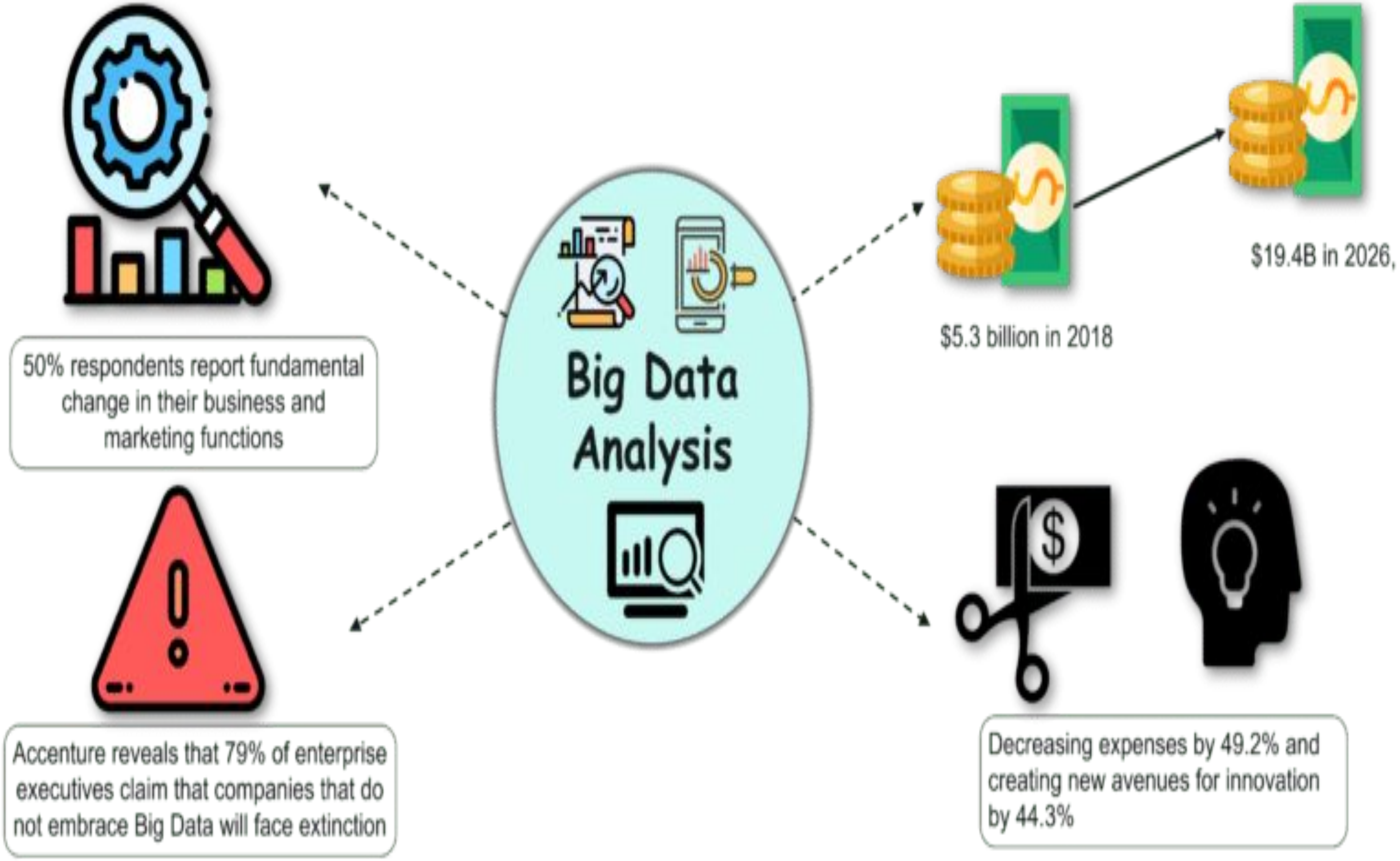
P&G

Procter&Gamble

Market Basket Analysis, analyses customer buying habits by finding associations between the different items that customers place in their “shopping baskets”

The company uses simulation models and predictive analysis in order to create the best design for its products.

Here are some Facts and Statistics by Forbes:



Big Data Analysis

50% respondents report fundamental change in their business and marketing functions

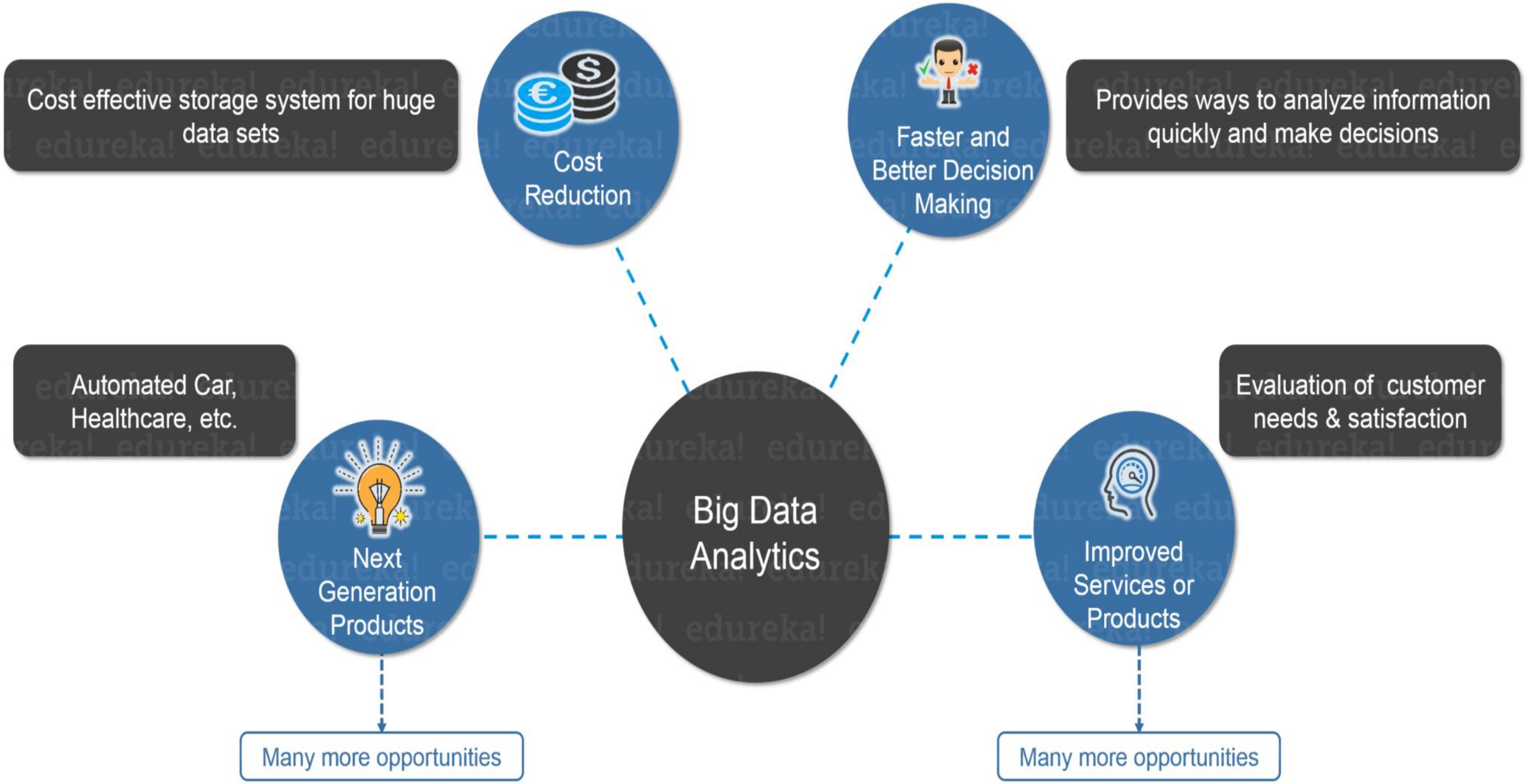
Accenture reveals that 79% of enterprise executives claim that companies that do not embrace Big Data will face extinction

$5.3 billion in 2018

$19.4B in 2026,

Decreasing expenses by 49.2% and creating new avenues for innovation by 44.3%

Big Data as an Opportunity



Cost effective storage system for huge data sets

Cost Reduction

Faster and Better Decision Making

Provides ways to analyze information quickly and make decisions

Automated Car, Healthcare, etc.

Next Generation Products

Big Data Analytics

Improved Services or Products

Evaluation of customer needs & satisfaction

Many more opportunities

Many more opportunities

Challenges in Big Data

- Big data consists of huge amount of data sets. The main challenge evolves in identifying the appropriate data from such mass of data and determining how to make best use of the relevant data.
- Even though the data and analysis method are determined, there is struggle in finding the appropriate and skilled manpower capable of working with both new technology and data analysis for relevant business insight.
- The variety of data types and formats may generate hindrance in the data analysis as it is very difficult to connect variety of data points for a single insight. Data integration is the important aspect of effective big data analysis, but it is also one of the major challenges that prevails.

Challenges in Big Data

- The technology landscape in the data world is evolving very fast. So, efficient handling of the technology along with adaptation to cope with technology is must for big data analysis.
- The organizational structure for big data project management should be apart from another project management task because this field is very much different and needs a strong and motivational project management team.
- During the big data analysis, the data analysts do not get full benefits from the data they have due to the security concerns about data protection.
- The technology infrastructure necessary to work with big data is very expensive.

- To overcome these challenges, a variety of technologies and approaches have been developed, including:
  - Distributed computing: Technologies such as Hadoop, Spark and Cloud-based data processing platforms allow for the distributed processing of large datasets, making it possible to analyze big data on commodity hardware.
  - NoSQL databases: These databases are designed to handle unstructured and semi-structured data, making it possible to store and process big data in its native form.
  - Stream processing: Technologies such as Apache Kafka, Apache Storm and Cloud-based streaming platforms allow for the real-time processing of big data streams, making it possible to analyze data as it is generated.
  - Machine learning: Machine learning algorithms can be used to analyze big data and extract insights from it, even when the data is of low quality.
  - Data governance: Implementing data governance strategies, such as data cataloging, lineage tracking, and access control, can help ensure the security, privacy, and compliance of big data.

Big Data Analytics

- Big Data analytics is a process used to extract meaningful insights, such as hidden patterns, unknown correlations, market trends, and customer preferences.
- Big data analytics is the process of examining large amounts of data of a variety of types.
- The primary goal of big data analytics is to help companies make better business decisions.
- Analyze huge volumes of transaction data as well as other data sources that may be left untapped by conventional business intelligence (BI) programs.

Big Data Analytics

- Big data analytics can be done with the software tools commonly used as part of advanced analytics disciplines Such as **predictive analysis and data mining.**
- But the unstructured data sources used for big data analytics may not fit in traditional data warehouses.
- Traditional data warehouses may not be able to handle the processing demands posed by big data.
- The technologies associated with big data analytics include NoSQL databases, Hadoop and MapReduce.

Big Data Analytics

- Using big data analytics to get results include four main areas of concern:
  - Basic Analytics
  - Advanced Analytics
  - Operationalized Analytics
  - Monetized Analytics

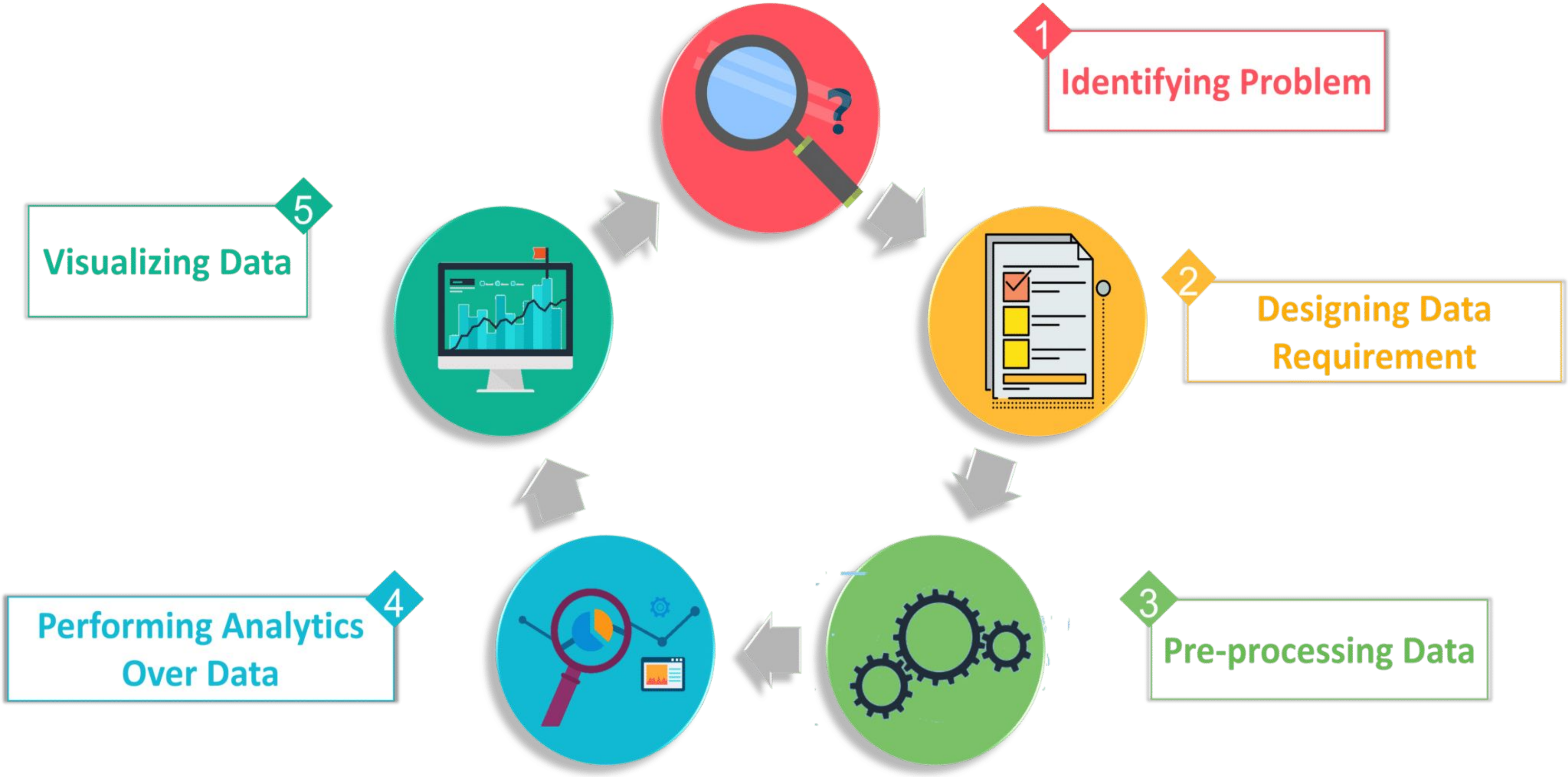
Big Data Analytics

- **Basic Analytics** : Can be used to explore your data, if you are not sure what you have, but you think something is of value
  - Slicing and Dicing
  - Basic Monitoring
  - Anomaly identification
- **Advanced analytics**: Provides algorithm for complex analysis of either structured or unstructured data.
  - Includes sophisticated statistical models, machine learning, neural networks, text analytics and other advanced data mining techniques
  - Can be deployed to find patterns in data, prediction, forecasting and complex event processing

Big Data Analytics

- **Operationalized Analytics**: It is an analysis of day to day operational data on a company.
  - Generally used in decision system of a company.
  - Example: Model for fraudulent claims in insurance company
- **Monetized Analytics**: Can be used to derive revenue above and beyond the insights it provides just for your own department or company.
  - You might be able to assemble a unique data set that is valuable to other companies as well.

Stages in Big Data Analytics



1 Identifying Problem

2 Designing Data Requirement

3 Pre-processing Data

4 Performing Analytics Over Data

5 Visualizing Data

Big Data Analytics Steps

- **Business case evaluation** - The Big Data analytics lifecycle begins with a business case, which defines the reason and goal behind the analysis.
- **Identification of data** - Here, a broad variety of data sources are identified.
- **Data filtering** - All of the identified data from the previous stage is filtered here to remove corrupt data.
- **Data extraction** - Data that is not compatible with the tool is extracted and then transformed into a compatible form.

Big Data Analytics Steps

- **Data aggregation** - In this stage, data with the same fields across different datasets are integrated.
- **Data analysis** - Data is evaluated using analytical and statistical tools to discover useful information.
- **Visualization of data** - With tools like Tableau, Power BI, and QlikView, Big Data analysts can produce graphic visualizations of the analysis.
- **Final analysis result** - This is the last step of the Big Data analytics lifecycle, where the final results of the analysis are made available to business stakeholders who will take action.

Types of Big Data Analytics

- **Descriptive Analytics**
  - Answers the question : “What happened? or is currently happening.”
- **Diagnostic Analytics**
  - Answers the question : “Why did this happen?”
- **Predictive Analytics**
  - Answers the question : “What might happen in the future?”
- **Prescriptive Analytics**
  - Answers the question :“What should we do next?”

Descriptive Analytics

- This summarizes past data into a form that people can easily read. This helps in creating reports, like a company's revenue, profit, sales, and so on. Also, it helps in the tabulation of social media metrics.
- Descriptive analytics answers the question, "What happened? or is currently happening."
- Use Case: The Dow Chemical Company analyzed its past data to increase facility utilization across its office and lab space. Using descriptive analytics, Dow was able to identify underutilized space. This space consolidation helped the company save nearly US $4 million annually.

Diagnostic Analytics

- This is done to understand what caused a problem in the first place. Techniques like drill-down, data mining, and data recovery are all examples. Organizations use diagnostic analytics because they provide an in-depth insight into a particular problem.
- Diagnostic analytics addresses the next logical question, “Why did this happen?”
- Use Case: An e-commerce company’s report shows that their sales have gone down, although customers are adding products to their carts. This can be due to various reasons like the form didn’t load correctly, the shipping fee is too high, or there are not enough payment options available. This is where you can use diagnostic analytics to find the reason.

Predictive Analytics

- **Predictive Analytics** :This type of analytics looks into the historical and present data to make predictions of the future. Predictive analytics uses data mining, AI, and machine learning to analyze current data and make predictions about the future. It works on predicting customer trends, market trends, and so on.
- Predictive analytics is used to make predictions about future trends or events and answers the question, “What might happen in the future?”
- Use Case: PayPal determines what kind of precautions they have to take to protect their clients against fraudulent transactions. Using predictive analytic, the company uses all the historical payment data and user behavior data and builds an algorithm that predicts fraudulent activities.

Prescriptive Analytics

- This type of analytics prescribes the solution to a particular problem. Prescriptive analytics works with both descriptive and predictive analytics. Most of the time, it relies on AI and machine learning.
- Prescriptive analytics answers the question, "What should we do next?"
- Use Case: Prescriptive analytics can be used to maximize an airline's profit. This type of analytics is used to build an algorithm that will automatically adjust the flight fares based on numerous factors, including customer demand, weather, destination, holiday seasons, and oil prices.

Role of Data Analyst

- A data analyst is someone who merely curates meaningful insights from data.
- A data analyst job roles involves looking at the known from new perspectives.
- A data analyst finds answers to a given set of questions from data.
- A data analyst addresses business problems but a data scientist not just addresses business problems but picks up those problems that will have the most business value once solved.
- Data analysts are the one who do the day-to-day analysis stuff

Data Analyst Skills

- Strong mathematical skills to help collect, measure, organize and analyze data
- Knowledge of programming languages like SQL,R, Python
- Technical proficiency regarding database design development, data models, techniques for data mining, and segmentation.
- Experience in handling reporting packages like Business Objects, ETL frameworks, databases
- Proficiency in statistics and statistical packages like Excel, SPSS, SAS to be used for data set analyzing
- Adept at using data processing platforms like Hadoop and Apache Spark
- Knowledge of data visualization software like Tableau, Qlik

Role of Data Scientist

- A data scientist is someone who can predict the future based on past patterns
- A data scientist job roles involves estimating the unknown.
- Data mining or extracting usable data from valuable data sources
- Using machine learning tools to select features, create and optimize classifiers
- Carrying out preprocessing of structured and unstructured data
- Enhancing data collection procedures to include all relevant information for developing analytic systems

Role of Data Scientist

- Processing, cleansing, and validating the integrity of data to be used for analysis
- Analyzing large amounts of information to find patterns and solutions
- Developing prediction systems and machine learning algorithms
- Presenting results in a clear manner
- Propose solutions and strategies to tackle business challenges
- Collaborate with Business and IT teams

Data Scientist Skills

- Programming Skills – knowledge of statistical programming languages like R, Python, and database query languages like SQL, Hive, Pig is desirable.
- Statistics – Good applied statistical skills, including knowledge of statistical tests, distributions, regression, maximum likelihood estimators, etc.
- Machine Learning – good knowledge of machine learning methods like k-Nearest Neighbors, Naive Bayes, SVM, Decision Forests.
- Strong Math Skills (Multivariable Calculus and Linear Algebra)
- Data Wrangling – proficiency in handling imperfections in data is an important aspect of a data scientist job description.
- Experience with Data Visualization Tools like Tableau , Power BI that help to visually encode data

Role of Distributed System in Big Data

- Big data consists of massive amount of data which cannot be stored in a single computer or node. - So, there is necessity for big data to be distributed across multiple nodes.
- Distributed system helps to solve big data problems without the requirement of a single resource capable to handle it. This makes the big data analytics cost efficient and performance improvement.
- Distributed systems are necessary for big data for several reasons:
- **Scalability:**
  - Big data is characterized by massive volumes of information that cannot be efficiently processed by a single machine. Distributed systems enable the horizontal scaling of resources by distributing the data and workload across multiple nodes, allowing systems to handle large and growing datasets.

Role of Distributed System in Big Data

- Distributed systems are necessary for big data for several reasons:
- **Parallel Processing:**
  - Distributed systems facilitate parallel processing, where tasks are divided among multiple nodes, and computations can be performed simultaneously. This parallelism significantly speeds up data processing and analysis, making it feasible to handle the vast amounts of data in big data scenarios.
- **Fault Tolerance:**
  - Big data systems often deal with hardware failures due to the sheer number of components involved. Distributed systems are designed to be fault-tolerant, ensuring that the failure of one or more nodes does not lead to the loss of data or system downtime.
- **Data Distribution and Replication:**
  - Big data is typically distributed across multiple storage nodes. Distributed systems provide mechanisms for efficient data distribution and replication, ensuring that data is available even if some nodes fail. This improves data availability and reliability.

Role of Distributed System in Big Data

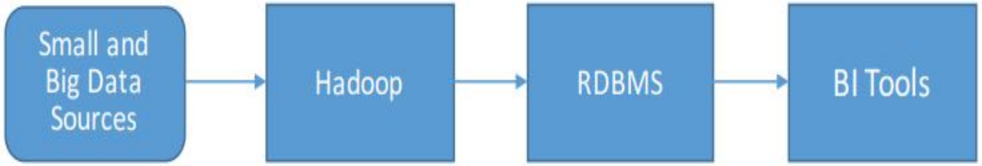
- Distributed systems are necessary for big data for several reasons:
- **Resource Utilization:**
  - Distributed systems allow for the efficient use of resources across multiple machines. By distributing the workload, each node can contribute to processing tasks, making the overall system more resource-efficient.
- **Data Locality:**
  - In big data scenarios, the data is often distributed across different nodes or storage clusters. Distributed systems support data locality, where computations can be performed near the data, minimizing data transfer times and improving overall system performance.
- **Global Accessibility:**
  - Big data may be generated, processed, and consumed globally. Distributed systems enable collaboration and access to data and resources from different geographical locations, supporting global enterprises and research efforts.
- **Adaptability to Heterogeneous Environments:**
  - Distributed systems can be deployed in heterogeneous environments with diverse hardware and software configurations. This adaptability is crucial for accommodating the variety of data sources and processing requirements in big data applications.

Role of Distributed System in Big Data

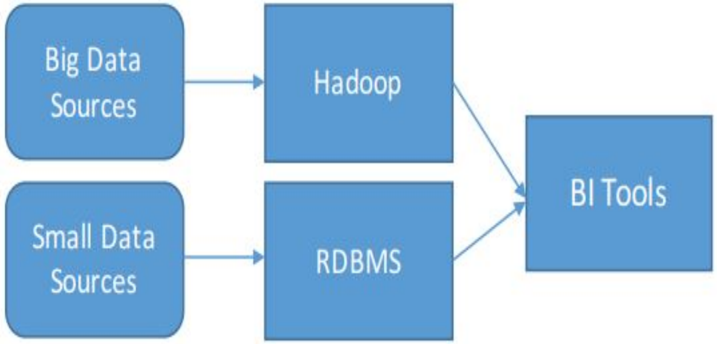
- Distributed systems are necessary for big data for several reasons:
- **Complex Analytics and Machine Learning:**
  - Big data often involves complex analytics, machine learning, and data processing tasks. Distributed systems, such as Apache Hadoop and Apache Spark, provide frameworks and tools specifically designed for these advanced analytics scenarios.
- **Cost-Effectiveness:**
  - Distributing the computational workload across multiple machines in a distributed system can be more cost-effective than investing in a single high-performance machine. This is particularly important for managing the costs associated with processing and storing large datasets.
- In summary, distributed systems provide the necessary infrastructure to address the challenges associated with big data, including scalability, fault tolerance, efficient resource utilization, and the ability to handle complex analytics. They enable organizations to effectively manage, process, and derive insights from massive datasets.

Big Data Warehouse systems

Hybrid approaches

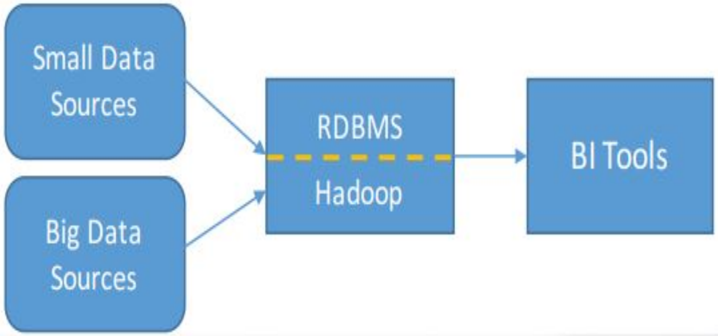


(A) Hadoop is used only for data ingestion/staging



(B) Big Data are kept separate from structured data: Hadoop is used as data management platform in parallel to the RDBMS.

Both platforms are used in conjunction for presentation purposes.



(C) Hadoop enhances RDBMS as data ingestion/staging tool, but also as data management and data presentation platform.

The "best" of both technologies is exploited.

Data lakes VS Data warehouses

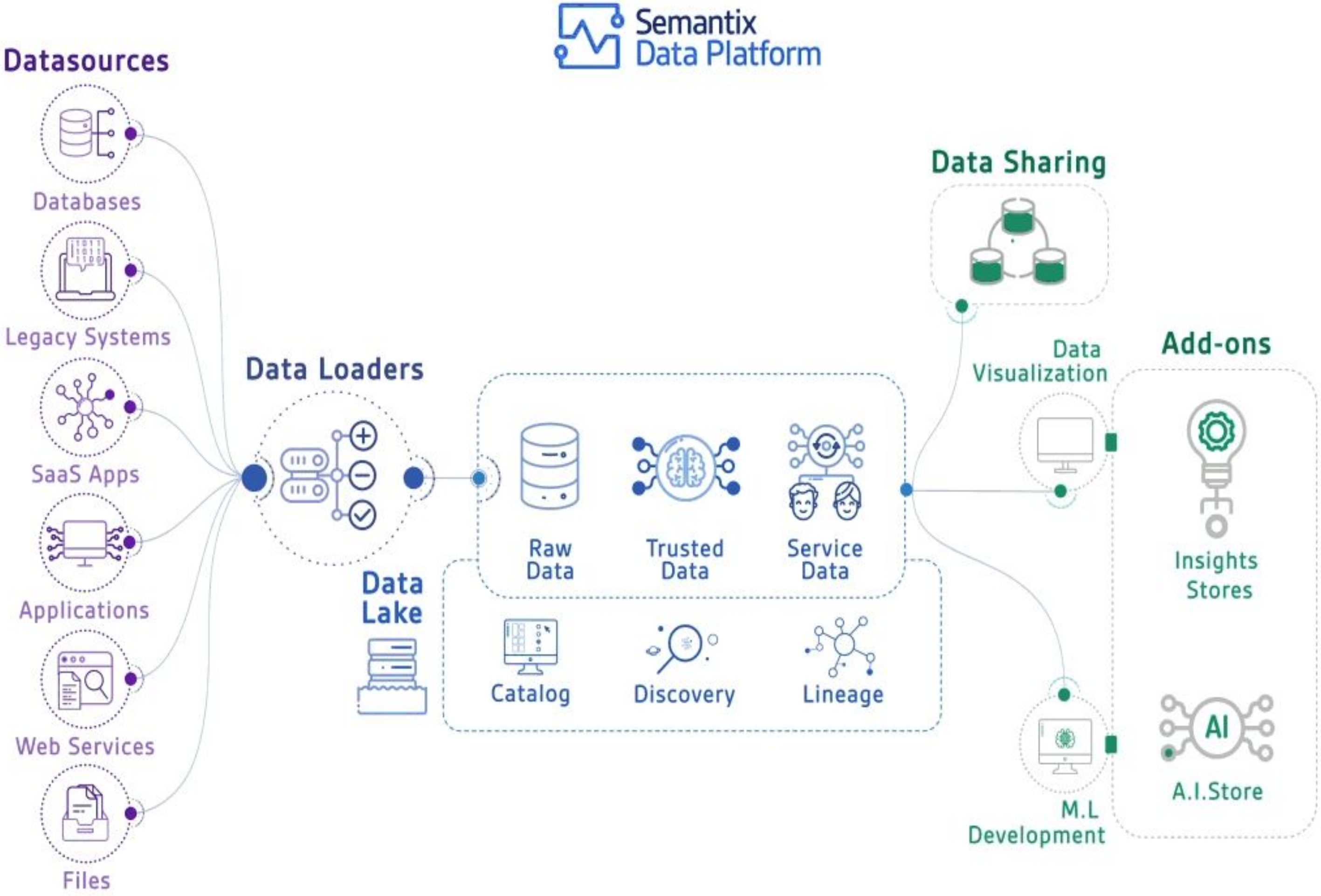
- Data lakes and data warehouses are both commonly used in big data technology, but they serve different purposes and have different characteristics.
- A data lake is a centralized repository that allows data to be stored in its raw, unstructured form. Data lakes are designed to handle a wide variety of data types and formats, including structured, semi-structured, and unstructured data. They are typically built on top of distributed file systems, such as Hadoop Distributed File System (HDFS), and can be accessed by a wide variety of tools and technologies. Data lakes are used for storing and processing large amounts of data in its raw format, providing a flexible and cost-effective way to store data for later use.

Data lakes VS Data warehouses

- A data warehouse, on the other hand, is a specialized, structured repository for storing and managing large amounts of data for reporting and analysis. Data warehouses are optimized for reading and querying large amounts of data and are typically built on top of relational databases, such as Oracle or SQL Server. They are designed to handle structured data and typically include a variety of tools and technologies for data modeling, ETL, and reporting. Data warehouses are used for reporting and analytics, providing a structured way to store data for later use.
- Data lakes are designed for storing and processing large amounts of raw data in its native format, while data warehouses are designed for storing and managing structured data for reporting and analysis. Data lakes are more flexible and cost-effective, while data warehouses are more optimized for reporting and analysis.

Current Trend in Big Data Analytics

- Data as service
- Responsible and Smarter Artificial Intelligence
- Predictive Analytics
- Quantum Computing
- Edge Computing
- Natural Language Processing
- Big data with cloud
- Dark Data
- Data Fabric
- Data Lakehouse
- Xops - XOps has emerged as the umbrella term for defining a combination of IT disciplines such as DevOps, DevSecOps, AIOps, MLOps, GitOps, and BizDevOps



Semantix Data Platform

Datasources

Databases

Legacy Systems

SaaS Apps

Applications

Web Services

Files

Data Loaders

Data Lake

Raw Data

Trusted Data

Service Data

Catalog

Discovery

Lineage

Data Sharing

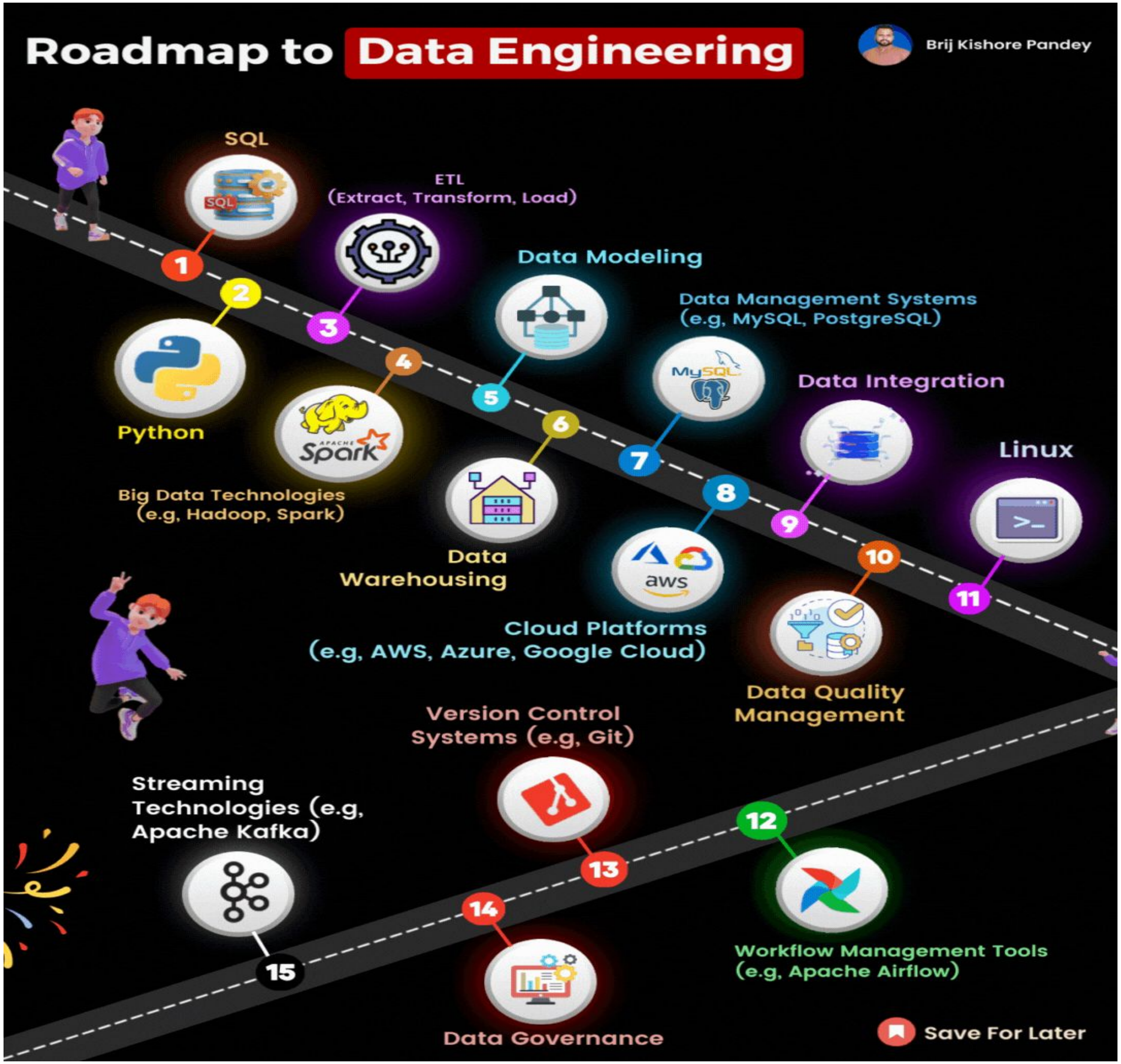
Data Visualization

M.L Development

Add-ons

Insights Stores

A.I.Store

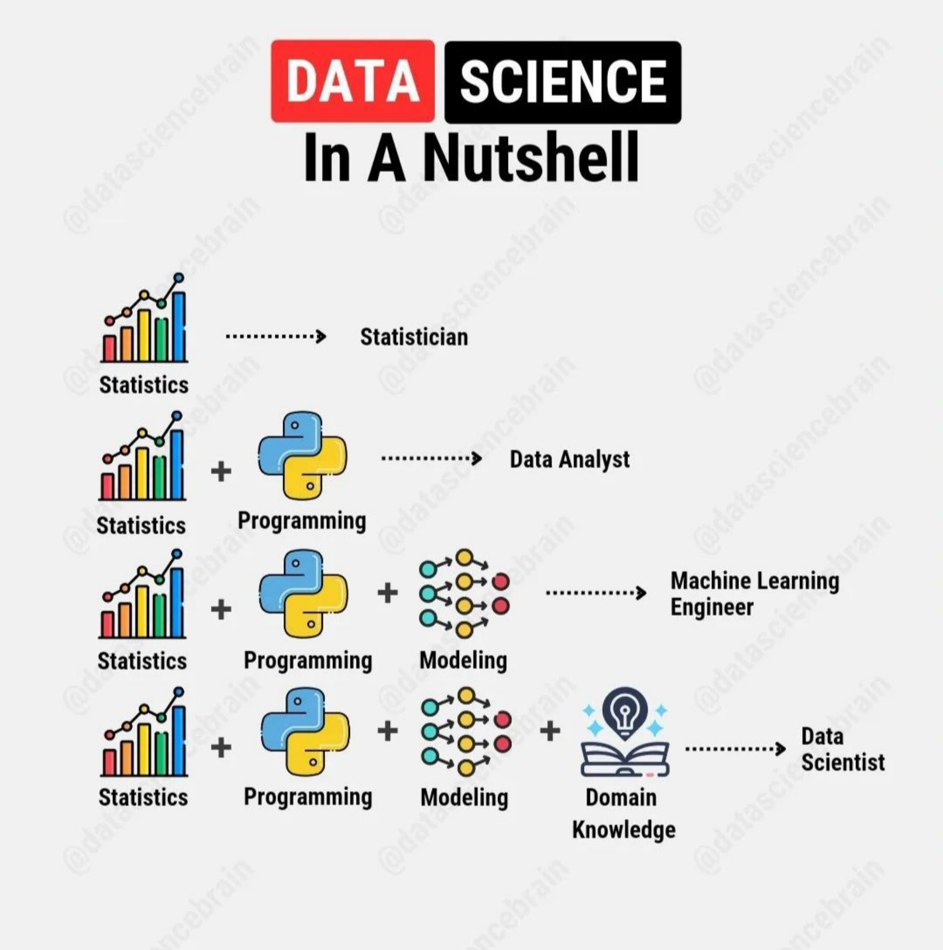


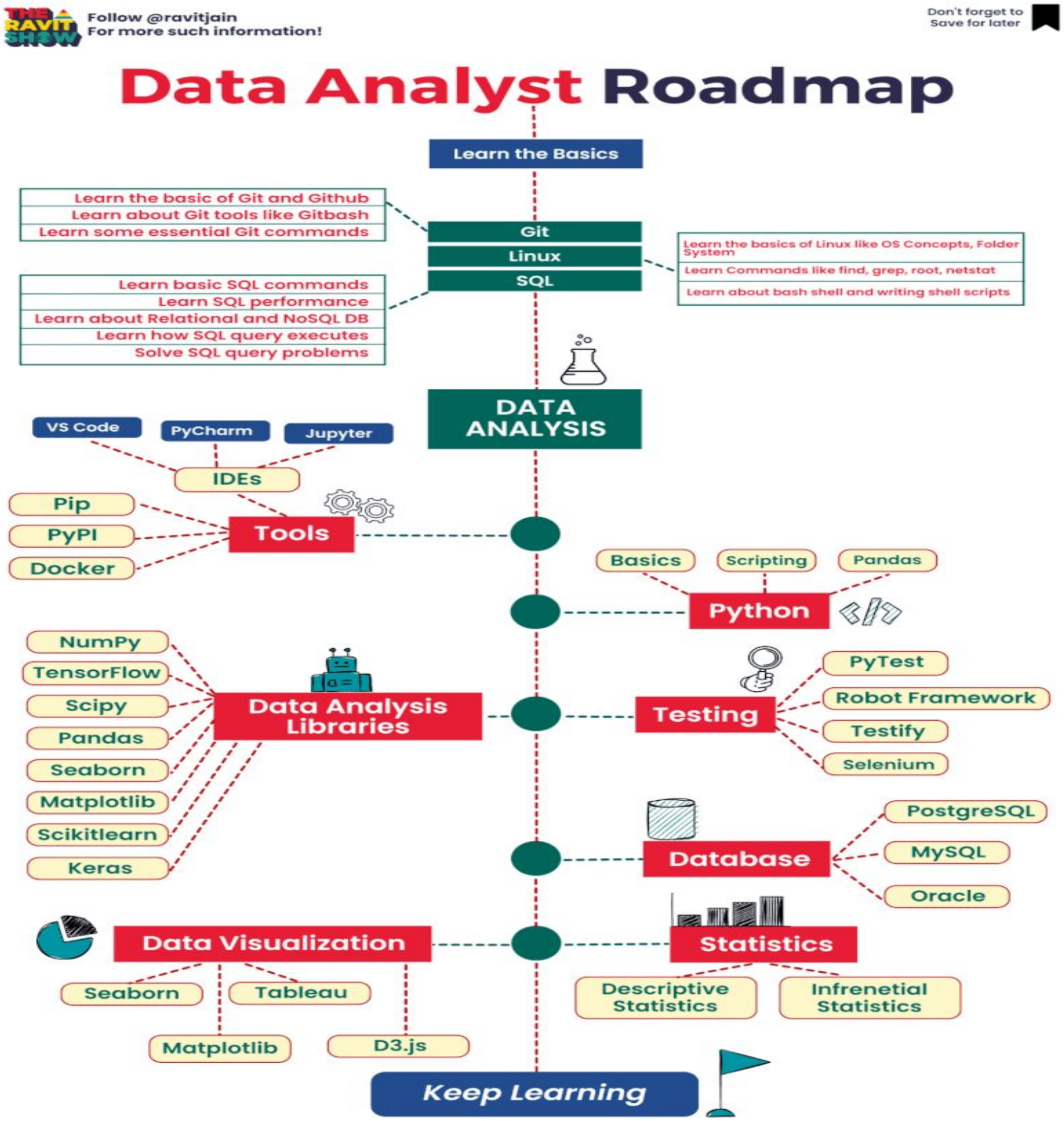
Roadmap to Data Engineering

Brij Kishore Pandey

1. SQL
2. Python
3. ETL (Extract, Transform, Load)
4. Big Data Technologies (e.g, Hadoop, Spark)
5. Data Modeling
6. Data Warehousing
7. Data Management Systems (e.g, MySQL, PostgreSQL)
8. Cloud Platforms (e.g, AWS, Azure, Google Cloud)
9. Data Integration
10. Data Quality Management
11. Linux
12. Workflow Management Tools (e.g, Apache Airflow)
13. Version Control Systems (e.g, Git)
14. Data Governance
15. Streaming Technologies (e.g, Apache Kafka)

Save For Later


DATA SCIENCE
In A Nutshell
Statistics
Statistician
Statistics
+
Programming
Data Analyst
Statistics
+
Programming
+
Modeling
Machine Learning Engineer
Statistics
+
Programming
+
Modeling
+
Domain Knowledge
Data Scientist



Follow @ravitjain
For more such information!

Don't forget to Save for later

Data Analyst Roadmap

Learn the Basics

Git
- Learn the basic of Git and Github
- Learn about Git tools like Gitbash
- Learn some essential Git commands

Linux
- Learn the basics of Linux like OS Concepts, Folder System
- Learn Commands like find, grep, root, netstat
- Learn about bash shell and writing shell scripts

SQL
- Learn basic SQL commands
- Learn SQL performance
- Learn about Relational and NoSQL DB
- Learn how SQL query executes
- Solve SQL query problems

DATA ANALYSIS

Tools
- IDEs: VS Code, PyCharm, Jupyter
- Pip
- PyPI
- Docker

Python
- Basics
- Scripting
- Pandas

Data Analysis Libraries
- NumPy
- TensorFlow
- Scipy
- Pandas
- Seaborn
- Matplotlib
- Scikitlearn
- Keras

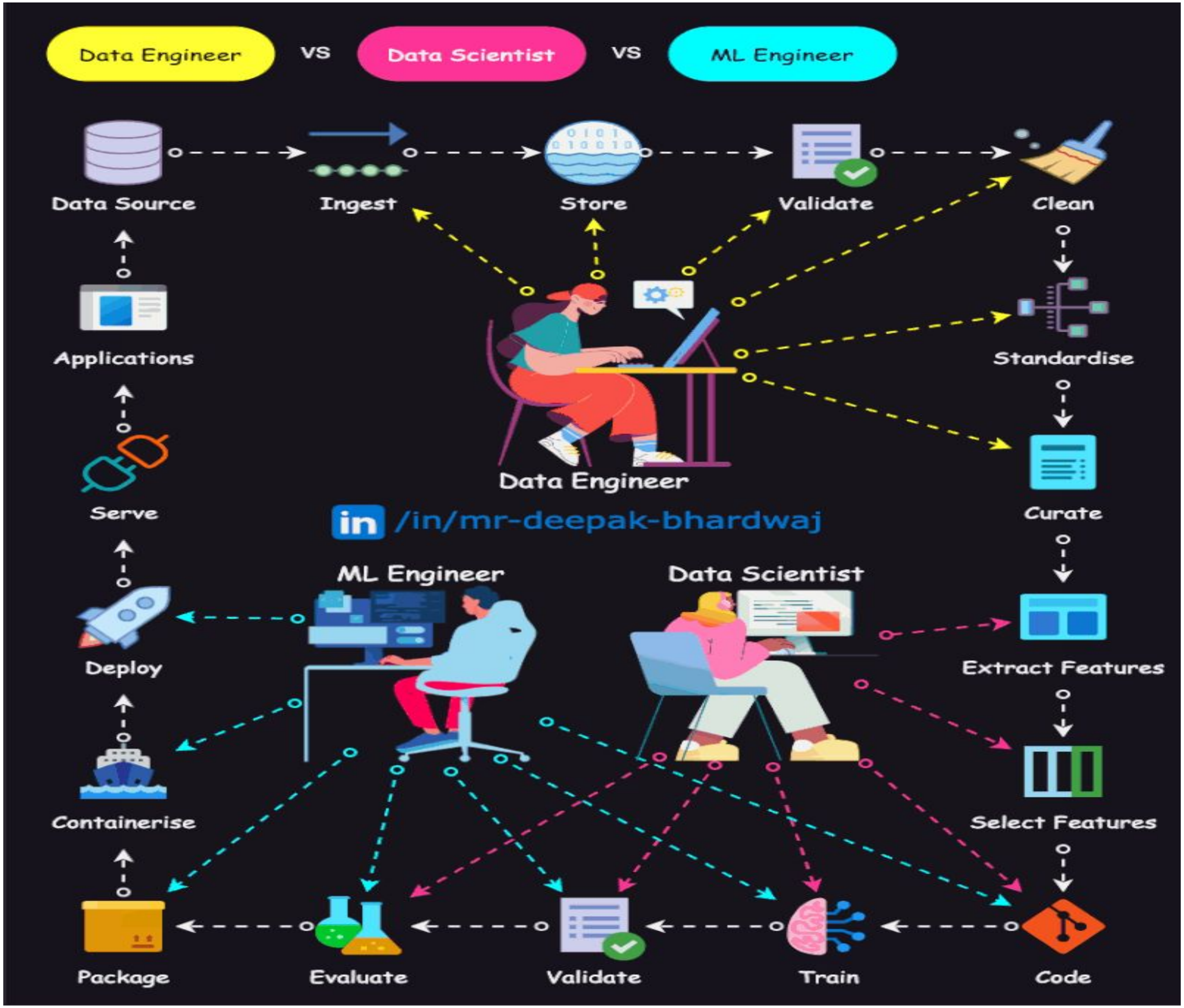
Testing
- PyTest
- Robot Framework
- Testify
- Selenium

Database
- PostgreSQL
- MySQL
- Oracle

Data Visualization
- Seaborn
- Tableau
- Matplotlib
- D3.js

Statistics
- Descriptive Statistics
- Infrenetial Statistics

Keep Learning


Data Engineer
vs
Data Scientist
vs
ML Engineer
Data Source
Ingest
Store
Validate
Clean
Applications
Standardise
Serve
Data Engineer
Curate
/in/mr-deepak-bhardwaj
ML Engineer
Data Scientist
Deploy
Extract Features
Containerise
Select Features
Package
Evaluate
Validate
Train
Code