Agenda

Big Data Analytics: Requirements, constraints, approaches, and technologies.

Big Data Systems – Characteristics: Failures; Reliability and Availability; Consistency – Notions of Consistency.

Contents

Terminologies Used in Big Data Environment

- In-Memory Analytics
- In-Database Processing
- Symmetric Multiprocessor System
- Massively Parallel Processing
- Difference between Parallel and Distributed Systems
- Shared Nothing Architecture
- Consistency, Availability, Partition Tolerance (CAP) Theorem Explained
- Basically Available Soft State Eventual Consistency (BASE)
- Few Top Analytics Tools

Picture This…

Scenario 1: You have heard a lot from your friends about the deals on offer on the Amazon site. You decide to register on www.amazon.co.in to avail their discount offers and bumper sales. A couple of days later, you make a purchase on their site. You landed yourself a good deal by going for books by your favorite author. There is something that does not escape your attention. Amazon has made a few suggestions (of books on similar topics or books by the same author) to you to help with your next or future purchases. You wonder how Amazon's recommendation engine was able to do this for you. Is it something that they do for all their customers? Well, Amazon's recommendation engine churns out these sort of good suggestions for customers like you, day in and day out. The company gathers all information about your past purchases together with what it knows about you, studies your buying patterns, and the buying patterns of customers like you to arrive at the recommendations that can help with your future purchase. At the core they have big data analytics working for them. 3.1

Scenario 2: You are the owner of a trucks transport company. Your company has 500 trucks plying several routes and carrying cargo from one place to another. It is one of those busy days where almost all the trucks are engaged in carrying cargo. You get a call to help with a cargo delivery. They are ready to pay double the charge. You do not want to miss this opportunity. But which truck should you engage. The one that is the nearest but is facing the heaviest traffic or the second nearest one but that is occupied to 75% and will not be able to take more load. There is a need to analyze the truck load, the fuel consumption, the traffic on various routes, etc. before deciding on which truck to select to pick up the new delivery.

3.1 Where do we Begin?

- Raw data is collected, classified, and organized.
- Associating it with adequate metadata and laying bare the context converts this data into meaningful information.
- It is then aggregated and summarized so that it becomes easy to consume it for analysis.
- Gradual accumulation of such meaningful information builds a knowledge repository.
- This, in turn, helps with actionable insights which prove useful for decision making. Refer Figure 3.1.
- Organizations have realized that they will not be able to ignore big data if they want to be competitive enough and make those timely decisions to make well of the fleeting opportunities.

They will have to analyze

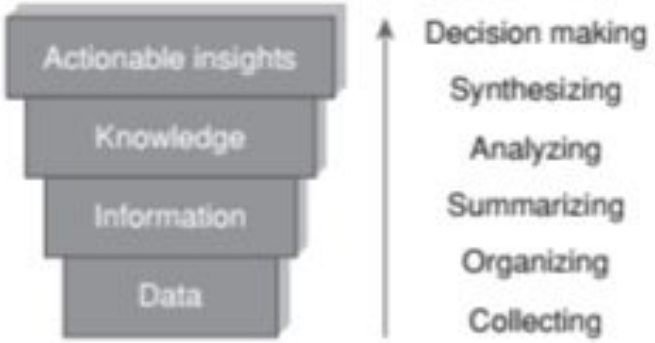

Figure 3.1 Transformation of data to yield actionable insights.

big time and also take into consideration big data that makes it to the organization at unprecedented level in terms of volume, velocity, and variety.

Big data analytics is the process of examining big data to uncover patterns, unearth trends, and find unknown correlations and other useful information to make faster and better decisions.

Analytics begin with analyzing all available data. Refer Figure 3.2

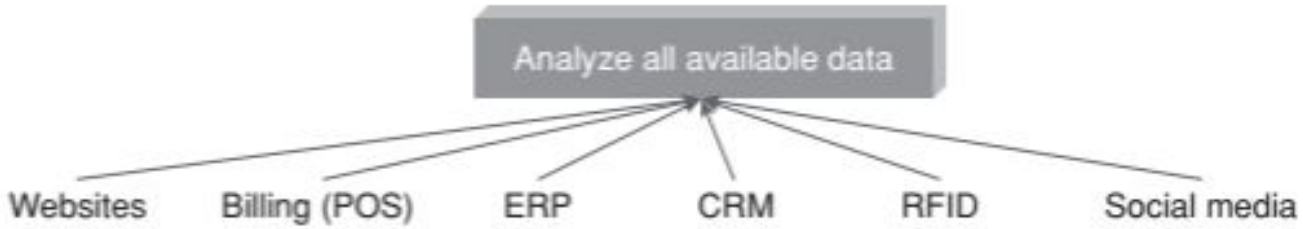

Figure 3.2 Types of unstructured data available for analysis.

3.2 What is Big Data Analytics? Big Data Analytics is…

1. Technology-enabled analytics: Quite a few data analytics and visualization tools are available in the market today from leading vendors such as IBM, Tableau, SAS, R Analytics, Statistica, World Programming Systems (WPS), etc. to help process and analyze your big data.
2. About gaining a meaningful, deeper, and richer insight into your business to steer it in the right direction, understanding the customer's demographics to cross-sell and up-sell to them, better leveraging the services of your vendors and suppliers, etc. Author's experience: The other day I was pleasantly surprised to get a few recommendations via email from one of my frequently visited online retailers. They had recommended clothing line from my favorite brand and also the color suggested was one to my liking. How did they arrive at this? In the recent past, I had been buying clothing line of a particular brand and the color preference was pastel shades. They had it stored in their database and pulled it out while making recommendations to me.

3. About a competitive edge over your competitors by enabling you with findings that allow quicker and better decision-making.

4. A tight handshake between three communities: IT, business users, and data scientists. Refer Figure 3.3.

5. Working with datasets whose volume and variety exceed the current storage and processing capabilities and infrastructure of your enterprise.

6. About moving code to data. This makes perfect sense as the program for distributed processing is tiny (just a few KBs) compared to the data (Terabytes or Petabytes today and likely to be Exabytes or Zettabytes in the near future).

What Big Data Analytics Isn't?

- We have often asked participants of our learning programs as what comes to mind when you hear the term "Big Data."
- And we are not surprised by the answer… it is "Volume."
- But now that we have a clear understanding of big data, we know it isn't only about volume but the variety and velocity too are very important factors.
- Figure 3.2 Types of unstructured data available for analysis.
- Websites Billing (POS) ERP CRM RFID Social media Analyze all available data

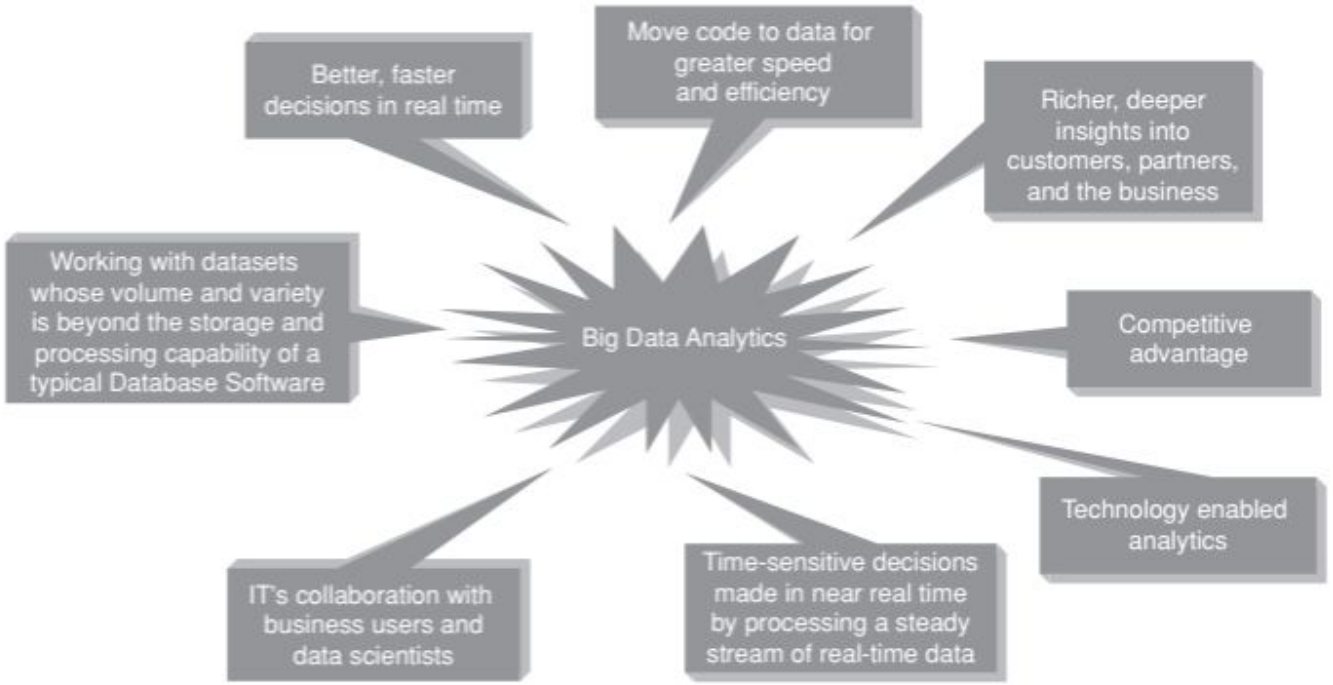

Figure 3.3 What is big data analytics?

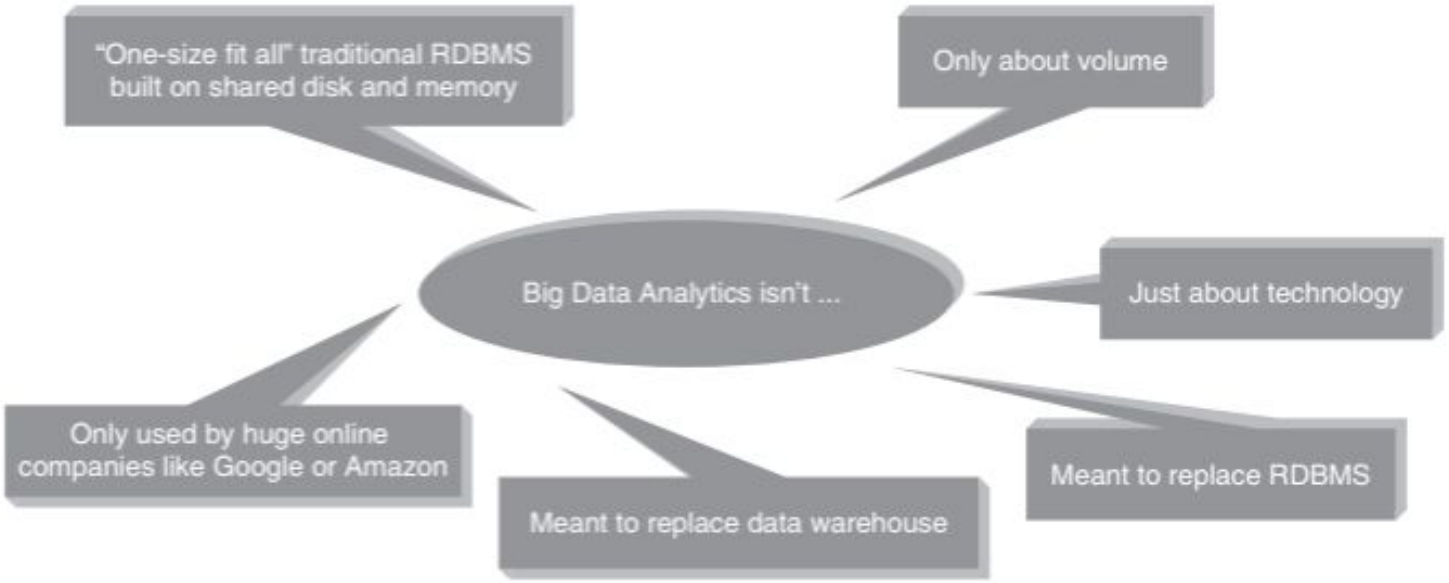

Figure 3.4 What big data analytics isn't?

Why this Sudden Hype Around Big Data Analytics?

If we go by the industry buzz, every place there seems to be talk about big data and big data analytics.

Why this sudden hype? Refer Figure 3.5.

Let us put it down to three foremost reasons:

1. Data is growing at a 40% compound annual rate, reaching nearly 45 ZB by 2020. In 2010, almost about 1.2 trillion Gigabyte of data was generated.
2. This amount doubled to 2.4 trillion Gigabyte in 2012 and to about 5 trillion Gigabytes in the year 2014. The volume of business data worldwide is expected to double every 1.2 years.

- Wal-Mart, the world retailer, processes one million customer transactions per hour.
- 500 million "tweets" are posted by Twitter users every day.
- 2.7 billion "Likes" and comments are posted by Facebook users in a day.
- Every day 2.5 quintillion bytes of data is created, with 90% of the world's data created in the past 2 years alone.

Source: (a) http://www.intel.com/content/www/us/en/communications/internet-minute-infographic.html

(b) http://www-01.ibm.com/software/data/bigdata/what-is-big-data.html

2. Cost per gigabyte of storage has hugely dropped.
3. There are an overwhelming number of user-friendly analytics tools available in the market today.

Classification of Analytics

There are basically two schools of thought:

1. Those that classify analytics into basic, operationalized, advanced, and monetized.
2. Those that classify analytics into analytics 1.0, analytics 2.0, and analytics 3.0.

Figure 3.5 What big data entails?

First School of Thought

1. Basic analytics: This primarily is slicing and dicing of data to help with basic business insights. This is about reporting on historical data, basic visualization, etc.
2. Operationalized analytics: It is operationalized analytics if it gets woven into the enterprise's business processes.
3. Advanced analytics: This largely is about forecasting for the future by way of predictive and prescriptive modeling.
4. Monetized analytics: This is analytics in use to derive direct business revenue.

3.5.2 Second School of Thought

Let us take a closer look at analytics 1.0, analytics 2.0, and analytics 3.0. Refer Table 3.1.

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.

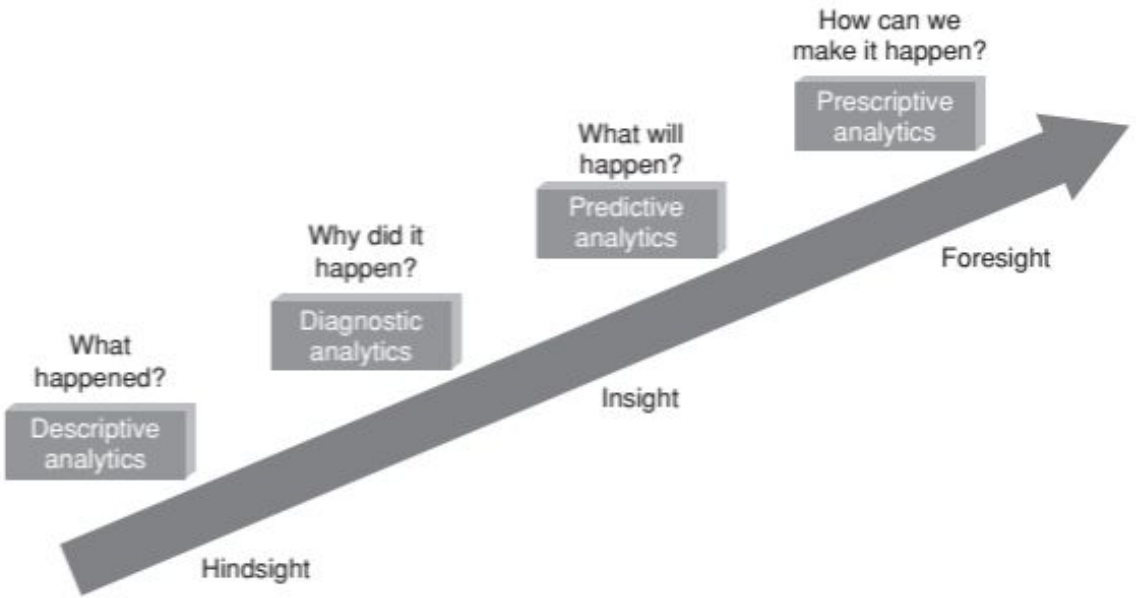

Figure 3.6 Analytics 1.0, 2.0, and 3.0.

Figure 3.6 shows the subtle growth of analytics from Descriptive → Diagnostic → Predictive → Prescriptive analytics.

Greatest Challenges that Prevent Businesses from Capitalizing on Big Data

1. Obtaining executive sponsorships for investments in big data and its related activities (such as training, etc.).
2. Getting the business units to share information across organizational silos.
3. Finding the right skills (business analysts and data scientists) that can manage large amounts of structured, semi-structured, and unstructured data and create insights from it.
4. Determining the approach to scale rapidly and elastically. In other words, the need to address the storage and processing of large volume, velocity, and variety of big data.
5. Deciding whether to use structured or unstructured, internal or external data to make business decisions.
6. Choosing the optimal way to report findings and analysis of big data (visual presentation and analytics) for the presentations to make the most sense.
7. Determining what to do with the insights created from big data.

Top Challenges Facing Big Data

1. Scale: Storage (RDBMS (Relational Database Management System) or NoSQL (Not only SQL)) is one major concern that needs to be addressed to handle the need for scaling rapidly and elastically. The need of the hour is a storage that can best withstand the onslaught of large volume, velocity, and variety of big data? Should you scale vertically or should you scale horizontally?

2. Security: Most of the NoSQL big data platforms have poor security mechanisms (lack of proper authentication and authorization mechanisms) when it comes to safeguarding big data. A spot that cannot be ignored given that big data carries credit card information, personal information, and other sensitive data.

3. Schema: Rigid schemas have no place. We want the technology to be able to fit our big data and not the other way around. The need of the hour is dynamic schema. Static (pre-defined schemas) are passé.

4. Continuous availability: The big question here is how to provide 24/7 support because almost all RDBMS and NoSQL big data platforms have a certain amount of downtime built in.

5. Consistency: Should one opt for consistency or eventual consistency?

6. Partition tolerant: How to build partition tolerant systems that can take care of both hardware and software failures?

7. Data quality: How to maintain data quality – data accuracy, completeness, timeliness, etc.? Do we have appropriate metadata in place?

3.8 Why is Big Data Analytics Important?

Let us study the various approaches to analysis of data and what it leads to.

1. Reactive – Business Intelligence: What does Business Intelligence (BI) help us with? It allows the businesses to make faster and better decisions by providing the right information to the right person at the right time in the right format. It is about analysis of the past or historical data and then displaying the findings of the analysis or reports in the form of enterprise dashboards, alerts, notifications, etc. It has support for both pre-specified reports as well as ad hoc querying.
2. 2. Reactive – Big Data Analytics: Here the analysis is done on huge datasets but the approach is still reactive as it is still based on static data.

3. Proactive – Analytics: This is to support futuristic decision making by the use of data mining, predictive modeling, text mining, and statistical analysis. This analysis is not on big data as it still uses the traditional database management practices on big data and therefore has severe limitations on the storage capacity and the processing capability.

4. Proactive – Big Data Analytics: This is sieving through terabytes, petabytes, exabytes of information to filter out the relevant data to analyze. This also includes high performance analytics to gain rapid insights from big data and the ability to solve complex problems using more data. 3.9 What Kind of Technologies are we Looking Toward

3.9 What Kind of Technologies are we Looking Toward to Help Meet the Challenges Posed by Big Data?

1. The first requirement is of cheap and abundant storage.
2. We need faster processors to help with quicker processing of big data.
3. Affordable open-source, distributed big data platforms, such as Hadoop.
4. Parallel processing, clustering, virtualization, large grid environments (to distribute processing to a number of machines), high connectivity, and high throughputs rather than low latency.
5. Cloud computing and other flexible resource allocation arrangements.

Data Science

Data science is the science of extracting knowledge from data. In other words, it is a science of drawing out hidden patterns amongst data using statistical and mathematical techniques.

It employs techniques and theories drawn from many fields from the broad areas of mathematics, statistics, information technology including machine learning, data engineering, probability models, statistical learning, pattern recognition and learning, etc.

Today we have a plethora of use-cases for “Data Science” that are already exploring massive datasets (Peta to Zetta bytes of Information) for weather predictions, oil drillings, seismic activities, financial frauds, terrorist network and activities, global economic impacts, sensor logs, social media analytics, and so many others beyond standard retail, manufacturing use-cases such as customer churn, market basket analytics (associative mining), collaborative filtering, regression analysis, etc. Data science is multi-disciplinary. Refer to Figure 3.7.

Business Acumen Skills

A data scientist should have the prowess to counter the pressures of business. A firm understanding of business domain further helps. The following is a list of traits that needs to be honed to play the role of data scientist.

1. Understanding of domain.
2. Business strategy.
3. Problem solving.
4. Communication.
5. Presentation.
6. Inquisitiveness.

Technology Expertise

It goes without saying that technology expertise will come in handy if one is to play the role of a data scientist. Cited below are few skills required as far as technical expertise is concerned.

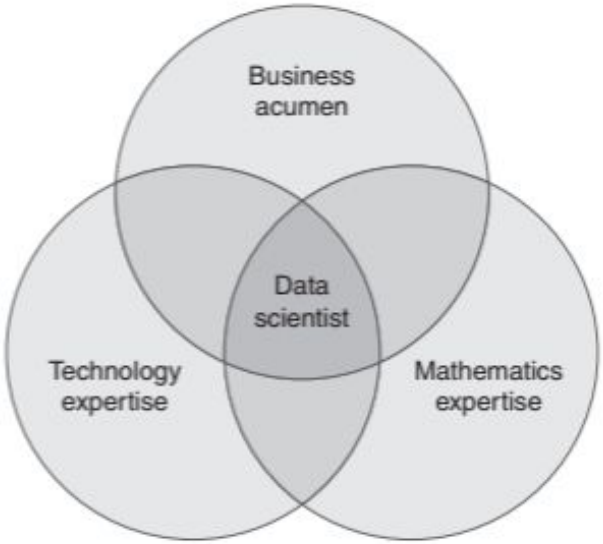

Figure 3.7 Data scientist.

1. Good database knowledge such as RDBMS.
2. 2. Good NoSQL database knowledge such as MongoDB, Cassandra, HBase, etc.
3. 3. Programming languages such as Java, Python, C++, etc.
4. 4. Open-source tools such as Hadoop.
5. 5. Data warehousing.
6. 6. Data mining.
7. 7. Visualization such as Tableau, Flare, Google visualization APIs, etc.

Mathematics Expertise

Since the core job of the data scientist will require him to comprehend data, interpret it, make sense of it, and analyze it, he/she will have to dabble in learning algorithms.

The following are the key skills that a data scientist will have to have in his arsenal.

1. Mathematics.
2. Statistics.
3. Artificial Intelligence (AI).
4. Algorithms.
5. Machine learning.
6. Pattern recognition.
7. Natural Language Processing.

To sum it up,

The data science process is

1. Collecting raw data from multiple disparate data sources.
2. Processing the data.
3. Integrating the data and preparing clean datasets.
4. 4. Engaging in explorative data analysis using model and algorithms.
5. Preparing presentations using data visualizations (commonly called Infographics, or BizAnalytics, or VizAnalytics, etc.)
6. Communicating the findings to all stakeholders.
7. Making faster and better decisions.

Data Scientist…Your New Best Friend!!!

In today's data age, a data scientist is the best friend that you can gift yourself. Refer Figure 3.8 to learn about the tasks that the data scientist can help you with.

3.11.1 Responsibilities of a Data Scientist Refer Figure 3.8.

1. Data Management: A data scientist employs several approaches to develop the relevant datasets for analysis. Raw data is just "RAW," unsuitable for analysis. The data scientist works on it to prepare it to reflect the relationships and contexts. This data then becomes useful for processing and further analysis.

2. Analytical Techniques: Depending on the business questions which we are trying to find answers to and the type of data available at hand, the data scientist employs a blend of analytical techniques to develop models and algorithms to understand the data, interpret relationships, spot trends, and unveil patterns.

3. Business Analysts: A data scientist is a business analyst who distinguishes cool facts from insights and is able to apply his business acumen and domain knowledge to see the results in the business context. He is a good presenter and communicator who is able to communicate the results of his findings in a language that is understood by the different business stakeholders.

Figure 3.8 Data scientist: your new best friend!!!