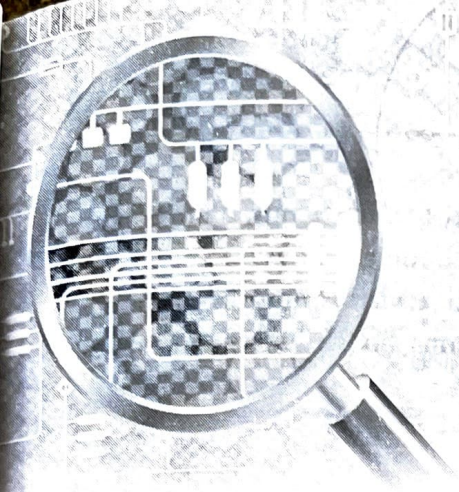


CHAPTER 3

Big Data Analytics

BRIEF CONTENTS

- What's in Store?
- Where do we Begin?
- What is Big Data Analytics?
- What Big Data Analytics isn't?
- Why this Sudden Hype Around Big Data Analytics?
- Classification of Analytics
- Greatest Challenges that Prevent Businesses from Capitalizing on Big Data
- Top Challenges Facing Big Data
- Why is Big Data Analytics Important?
- What Kind of Technologies are we Looking Toward to Help Meet the Challenges Posed by Big Data?
- Data Science
- Data Scientist … Your New Best Friend!!!
- 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

"If you do not know how to ask the right question, you discover nothing."

– W. Edwards Deming

WHAT'S IN STORE?

This chapter is about understanding "Big Data Analytics." We have taken you through the comprehension of the term Big Data – datasets which are voluminous, rich in variety, and calls for processing at a great speed. Big data analytics is the process of examining these large datasets of big data – to unearth hidden

patterns, decipher unknown correlations, understand the rationale behind market trends, and recognize customer preferences and other useful business information. The analytical findings can lead to more effective marketing, better customer service and satisfaction, newer products and services, improved operational efficiency, reduced expenditure, competitive advantages over rival organizations, boosted business gains, etc.

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.

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

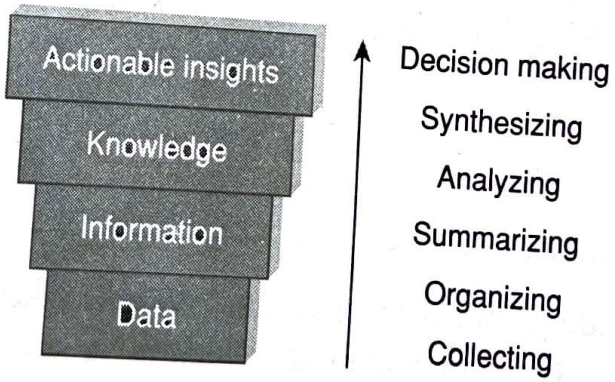


Figure 3.1 Transformation of data to yield actionable insights.

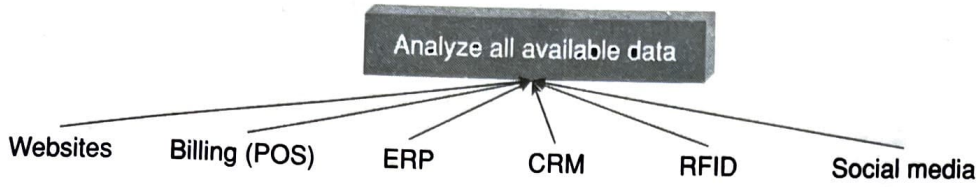


Figure 3.2 Types of unstructured data available for analysis.

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.

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).

3.3 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.

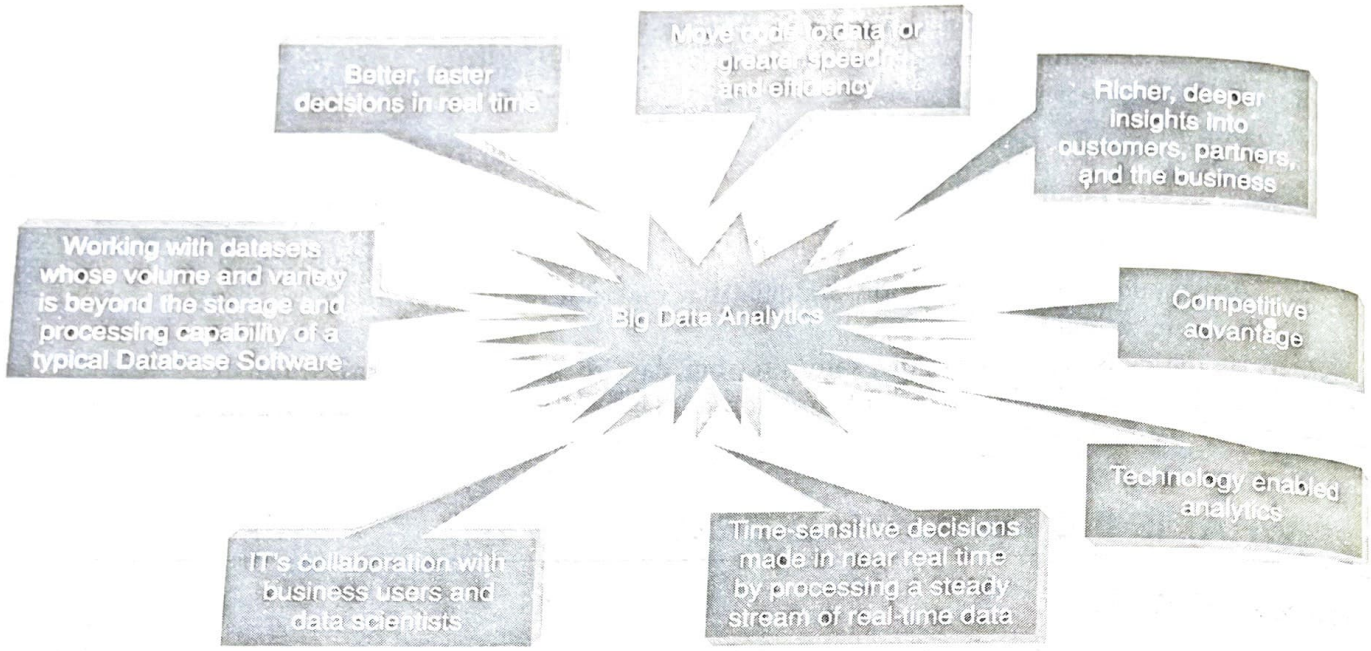


Figure 3.3 What is big data analytics?

Refer Figure 3.4. Big data isn't just about technology. It is about understanding what the data is saying to us. It is about understanding relationships that we thought never existed between datasets. It is about patterns and trends waiting to be unveiled.

And of course, big data analytics is not here to replace our now very robust and powerful Relational Database Management System (RDBMS) or our traditional Data Warehouse. It is here to coexist with both RDBMS and Data Warehouse, leveraging the power of each to yield business value. Big data analytics is not "One-size fits all" traditional RDBMS built on shared disk and memory.

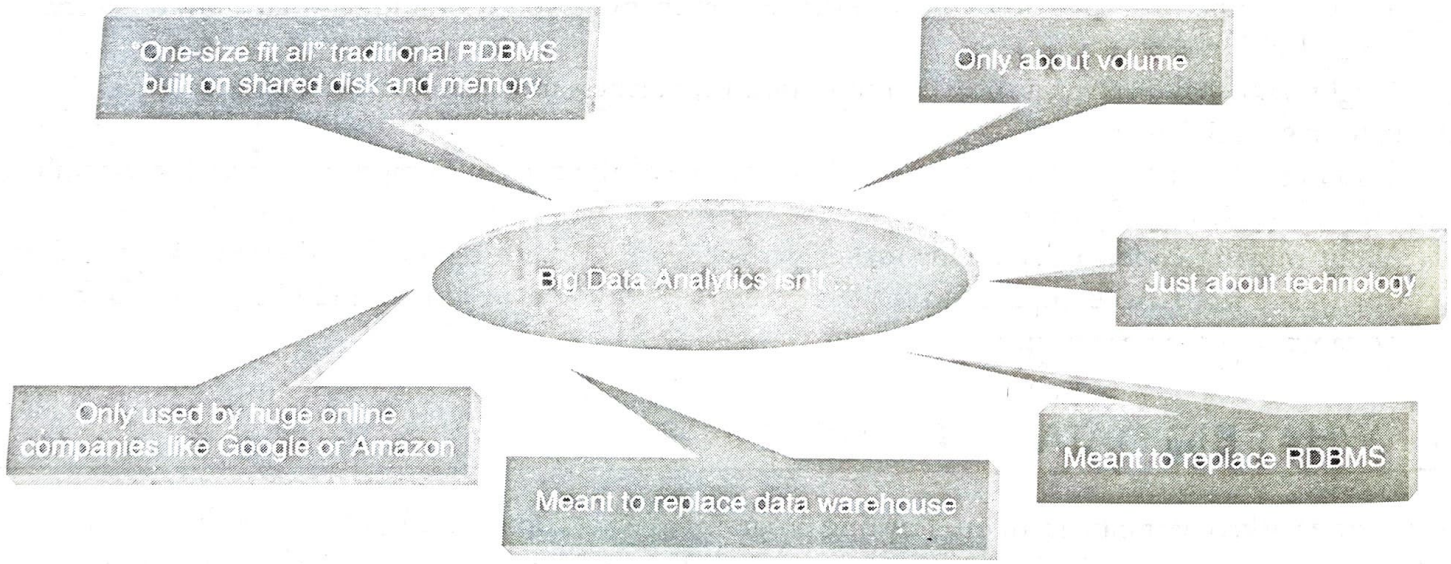


Figure 3.4 What big data analytics isn't?

And before we think it is only used by huge online companies like a Google or Amazon, let us clear the myth. It is for any business and any industry that needs actionable insights out of their data (both internal and external).

3.4 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. 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.

3.5 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.

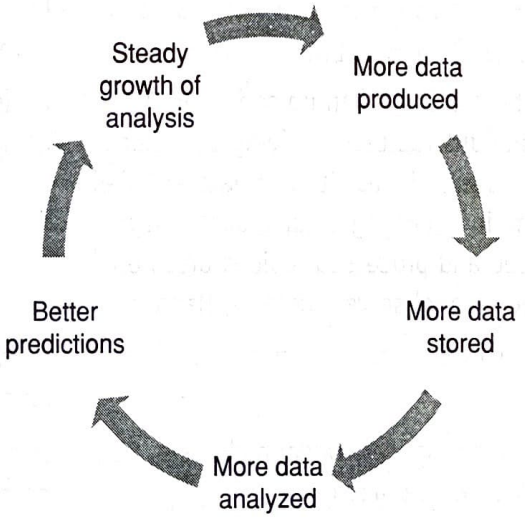


Figure 3.5 What big data entails?

3.5.1 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.

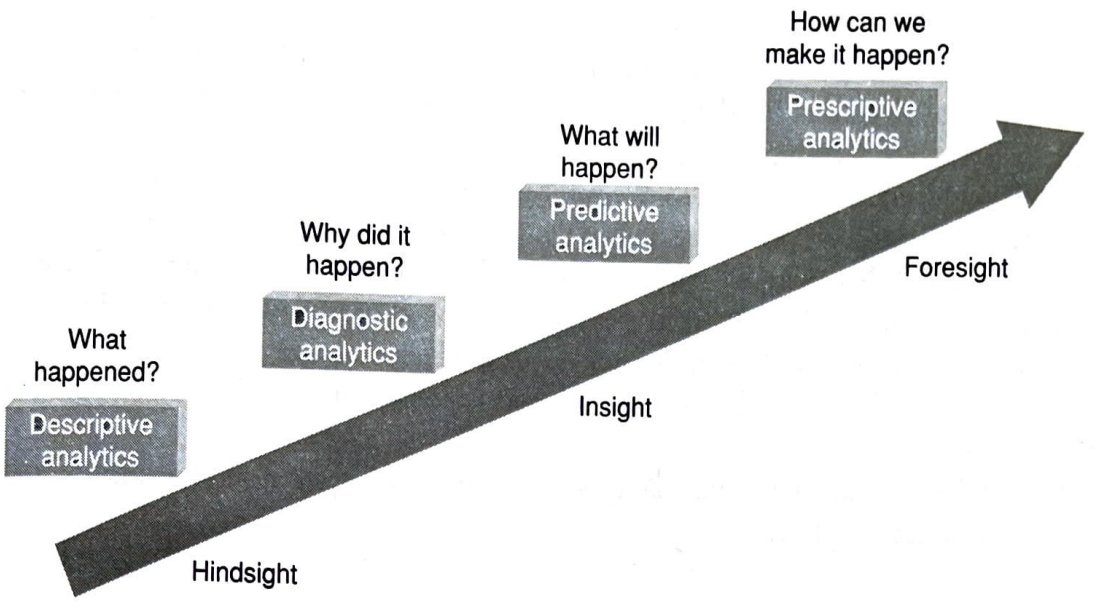


Figure 3.6 Analytics 1.0, 2.0, and 3.0.

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

3.6 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.

3.7 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. **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 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.

3.10 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.

3.10.1 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.

3.10.2 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.

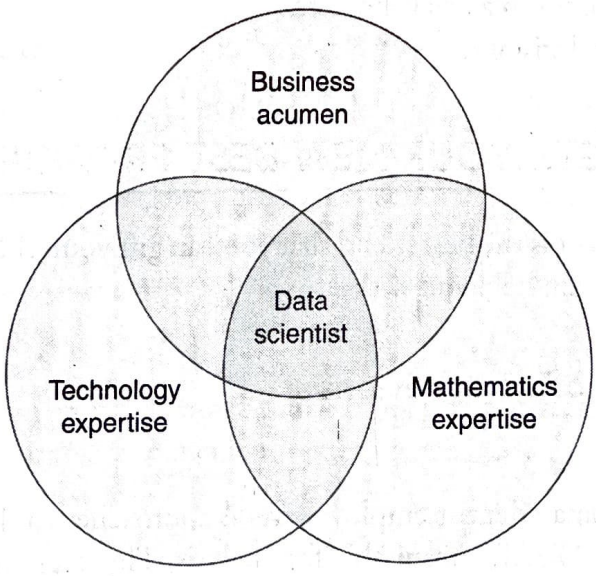


Figure 3.7 Data scientist.

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

3.10.3 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. 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.

3.11 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.

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

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.

3.12 TERMINOLOGIES USED IN BIG DATA ENVIRONMENTS

In order to get a good handle on the big data environment, let us get familiar with a few key terminologies in this arena.

3.12.1 In-Memory Analytics

Data access from non-volatile storage such as hard disk is a slow process. The more the data is required to be fetched from hard disk or secondary storage, the slower the process gets. One way to combat this challenge is to pre-process and store data (cubes, aggregate tables, query sets, etc.) so that the CPU has to fetch a small subset of records. But this requires thinking in advance as to what data will be required for analysis. If there is a need for different or more data, it is back to the initial process of pre-computing and storing data or fetching it from secondary storage.

This problem has been addressed using in-memory analytics. Here all the relevant data is stored in Random Access Memory (RAM) or primary storage thus eliminating the need to access the data from hard disk. The advantage is faster access, rapid deployment, better insights, and minimal IT involvement.

3.12.2 In-Database Processing

In-database processing is also called as *in-database analytics*. It works by fusing data warehouses with analytical systems. Typically the data from various enterprise On Line Transaction Processing (OLTP) systems after

cleaning up (de-duplication, scrubbing, etc.) through the process of ETL is stored in the Enterprise Da Warehouse (EDW) or data marts. The huge datasets are then exported to analytical programs for comple and extensive computations. With in-database processing, the database program itself can run the compu tations eliminating the need for export and thereby saving on time. Leading database vendors are offerin this feature to large businesses.

3.12.3 Symmetric Multiprocessor System (SMP)

In SMP, there is a single common main memory that is shared by two or more identical processors. The processors have full access to all I/O devices and are controlled by a single operating system instance.

SMP are tightly coupled multiprocessor systems. Each processor has its own high-speed memory, called cache memory and are connected using a system bus. Refer Figure 3.9.

3.12.4 Massively Parallel Processing

Massive Parallel Processing (MPP) refers to the coordinated processing of programs by a number of processors working parallel. The processors, each have their own operating systems and dedicated memory. They work on different parts of the same program. The MPP processors communicate using some sort of messaging interface. The MPP systems are more difficult to program as the application must be divided in such a way that all the executing segments can communicate with each other. MPP is different from Symmetrically Multiprocessing (SMP) in that SMP works with the processors sharing the same operating system and same memory. SMP is also referred to as *tightly-coupled multiprocessing*.

3.12.5 Difference Between Parallel and Distributed Systems

The next two terms that we discuss are parallel and distributed systems.

As is evident from Figure 3.10, a parallel database system is a tightly coupled system. The processors co-operate for query processing. The user is unaware of the parallelism since he/she has no access to a specific

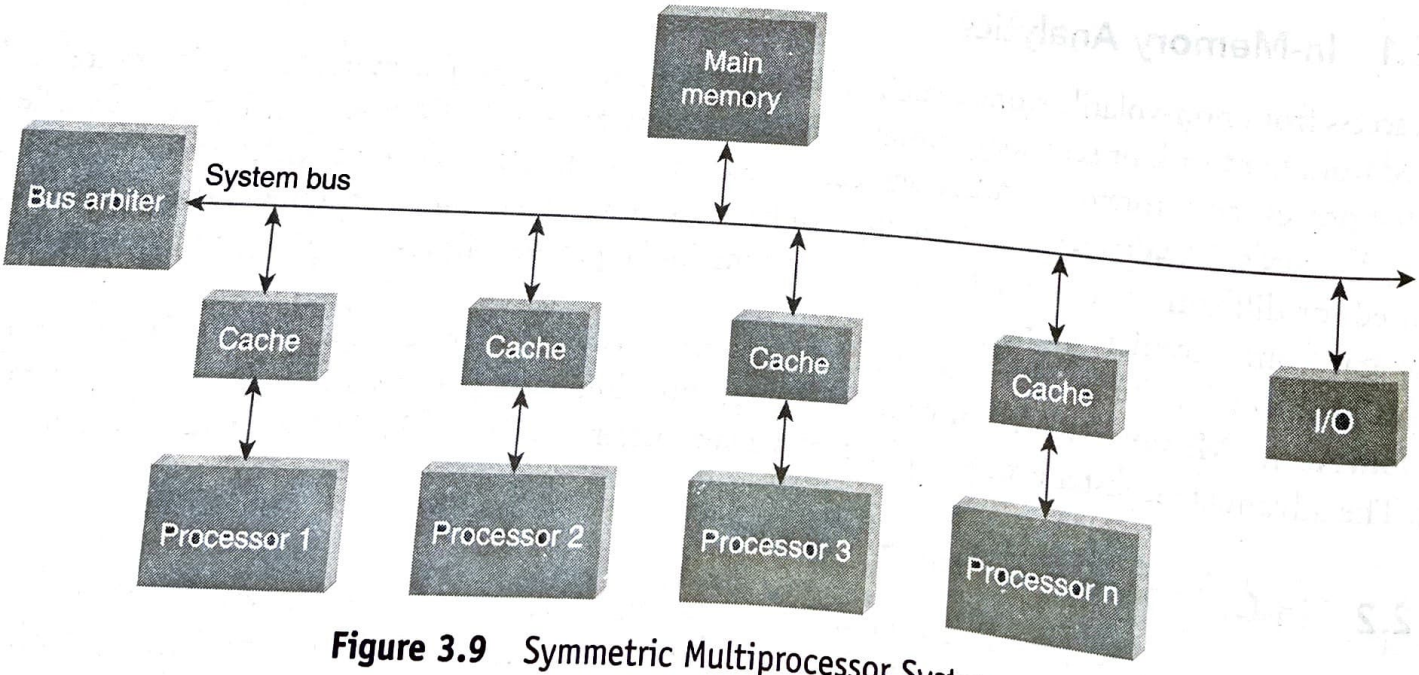


Figure 3.9 Symmetric Multiprocessor System.

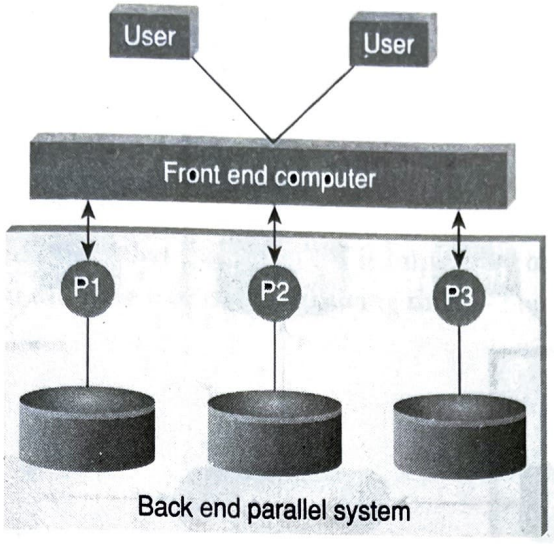


Figure 3.10 Parallel system.

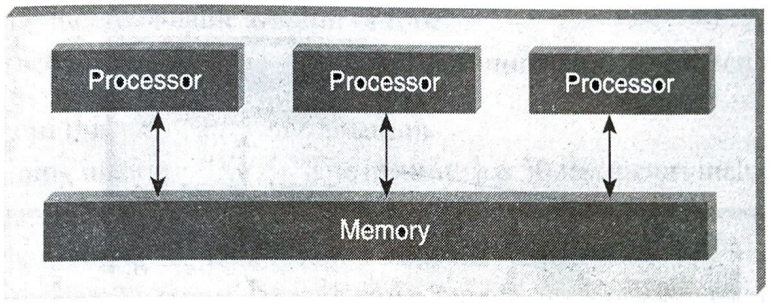


Figure 3.11 Parallel system.

processor of the system. Either the processors have access to a common memory (Refer Fig 3.11) or make use of message passing for communication.

Distributed database systems are known to be loosely coupled and are composed by individual machines. Refer Figure 3.12. Each of the machines can run their individual application and serve their own respective user. The data is usually distributed across several machines, thereby necessitating quite a number of machines to be accessed to answer a user query. Refer Figure 3.13.

3.12.6 Shared Nothing Architecture

Let us look at the three most common types of architecture for multiprocessor high transaction rate systems. They are:

1. Shared Memory (SM).
2. Shared Disk (SD).
3. Shared Nothing (SN).

In shared memory architecture, a common central memory is shared by multiple processors. In shared disk architecture, multiple processors share a common collection of disks while having their own private memory. In shared nothing architecture, neither memory nor disk is shared among multiple processors.

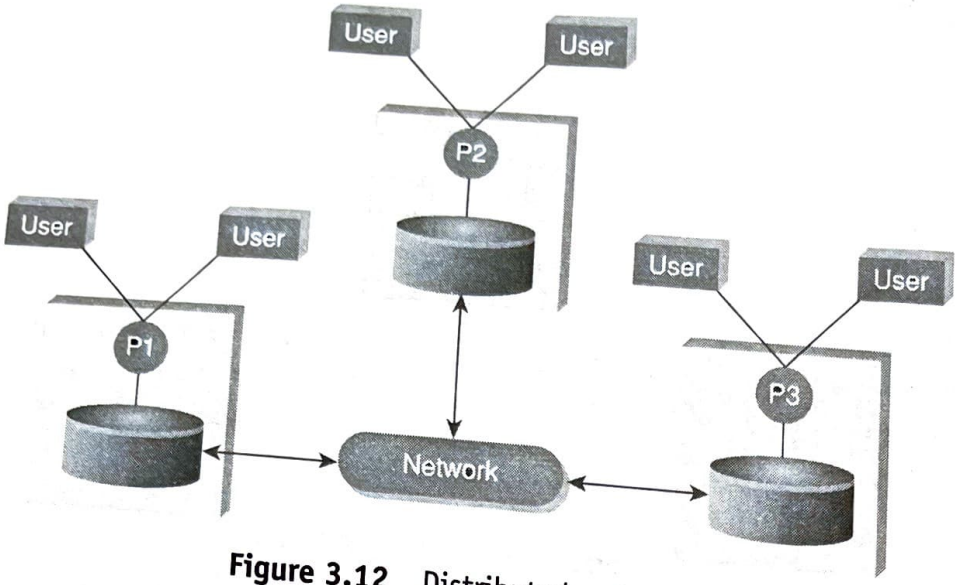


Figure 3.12 Distributed system.

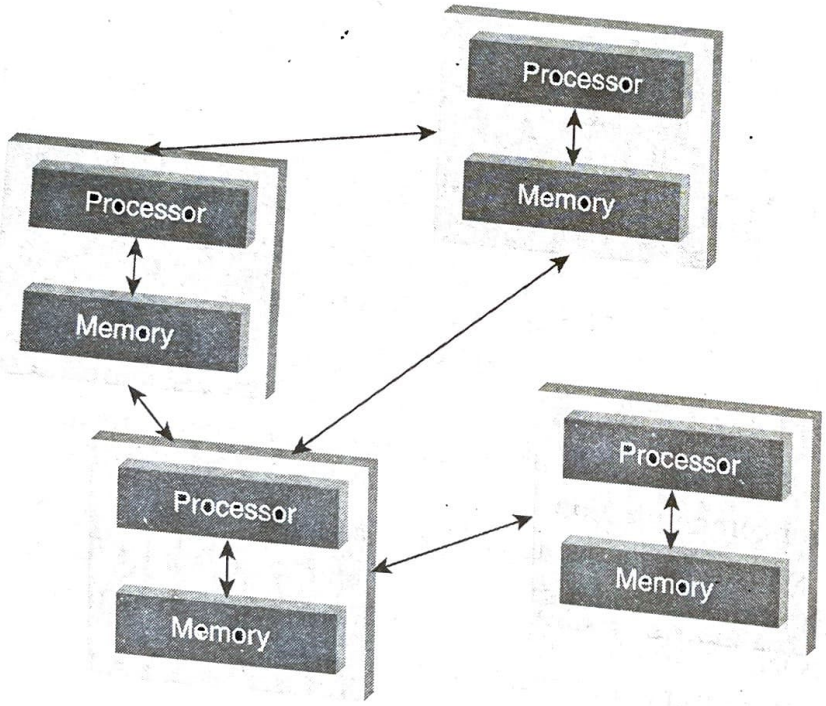


Figure 3.13 Distributed system.

3.12.6.1 Advantages of a "Shared Nothing Architecture"

1. **Fault Isolation:** A "Shared Nothing Architecture" provides the benefit of isolating fault. A fault in a single node is contained and confined to that node exclusively and exposed only through messages (or lack of it).
2. **Scalability:** Assume that the disk is a shared resource. It implies that the controller and the disk bandwidth are also shared. Synchronization will have to be implemented to maintain a consistent shared state. This would mean that different nodes will have to take turns to access the critical data. This

imposes a limit on how many nodes can be added to the distributed shared disk system, thus compromising on scalability.

3.12.7 CAP Theorem Explained

The CAP theorem is also called the *Brewer's Theorem*. It states that in a distributed computing environment (a collection of interconnected nodes that share data), it is impossible to provide the following guarantees. Refer Figure 3.14. At best you can have two of the following three – one must be sacrificed.

1. Consistency
2. Availability
3. Partition tolerance

3.12.7.1 CAP Theorem

Let us spend some time understanding the earlier mentioned terms.

1. Consistency implies that every read fetches the last write.
2. Availability implies that reads and writes always succeed. In other words, each non-failing node will return a response in a reasonable amount of time.
3. Partition tolerance implies that the system will continue to function when network partition occurs.

Let us try to understand this using a real-life situation.

You work for a training institute, "XYZ." The institute has 50 instructors including you. All of you report to a training coordinator. At the end of the month, all the instructors together with the training coordinator peruse through the training requests received from the various corporate houses and prepare a training schedule for each instructor. These training schedules (one for each instructor) are shared with "Amey," the office administrator. Each morning, you either call the office helpdesk (essentially Amey's desk) or check in-person with Amey for your schedule for the day. In case a training request has been cancelled or updated (updates can be in the form of change in course, change in duration, change of the training timings, etc.), Amey is informed of the updates and the schedules are subsequently updated by him.

Things were good until now. Few corporate houses were your clients and the schedules of each instructor could be smoothly managed without any major hiccups. But your training institute has been implementing promotion campaigns to expand the business. As a result of advertising in the media and word of mouth publicity by your existing clients, you suddenly see an upsurge in training requests from existing and new clients. In consequence of that, more instructors have been recruited. Few trainers/consultants have also been roped in from other training institutes to help tackle the load.

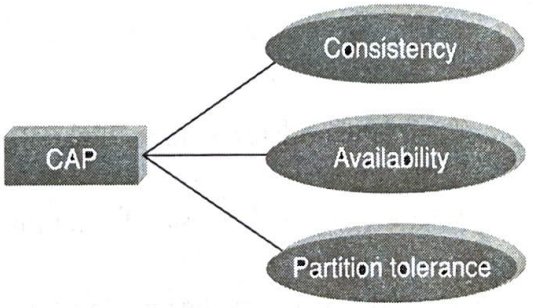


Figure 3.14 Brewer's CAP.

Now when you go to Amey to check your schedule or call in at the helpdesk, you are prepared for a wait in the queue. Looking at the current state of affairs, the training coordinator decides to recruit an additional office administrator "Joey." The helpdesk number will remain the same and will be shared by both the office administrators.

This arrangement works well for a couple of days. Then one day…

You: Hey Amey!

Amey: Hi! How can I help?

You: I think I am scheduled to anchor a training at 3:00 pm today. Can I please have the details?

Amey: Sure! Just a minute.

Amey browses through the file where he maintains the schedules. He does not see a training scheduled against your name at 3:00 pm today and responds back, "You do not have any training to conduct at 3:00 pm."

You: How is that possible? The training coordinator called up yesterday evening to inform of the same and said he has updated the office administrators of the same.

Amey: Oh! Did he say which office administrator? It could have been Joey. Please check with Joey.

Amey: Hey Joey! Please check the schedule for Paul here… Do you see something scheduled at 3:00 pm today?

Joey: Sure enough! He is anchoring the training for client "Z" today at 3:00 pm.

A clear case of inconsistent system!!! The updates in the schedule were shared by the training coordinator with Joey and you were checking for your schedule with Amey.

You share this incident with the training coordinator and that gets him thinking. The issue has to be addressed immediately otherwise it will be difficult to avoid a chaotic situation. He comes up with a plan and shares it with both the office administrators the following day.

Training Coordinator: Folks, each time that either an instructor or me calls any one of you to update a schedule, make sure that both of you update it in your respective files. This way the instructor will always get the most recent and consistent information irrespective of whom amongst the two of you he/she speaks to.

Joey: But that could mean a delay in answering either a phone call or sharing the schedule with the instructor waiting in queue.

Training Coordinator: Yes, I understand. But there is no way that we can give incorrect information.

Amey: There is this other problem as well. Suppose one of us is on leave on a particular day. That would mean that we cannot take any update related calls as we will not be able to simultaneously update both the files (my file and Joey's).

Training Coordinator: Well, good point! *That's the availability problem!!!* But I have thought about that as well. Here is the plan:

1. If one of you receives the update call (any updates to any schedule), ensure that you inform the other person if he is available.
2. In case the other person is not available, ensure that you inform him of all the updates to all schedules via email. It is a must!!!
3. When the other person resumes duty, the first thing he will do is update his file with all the updates to all schedules that he has received via email.

Wow!!! That is sure a Consistent and Available system!!!

Looks like everything is in control. Wait a minute! There is a tiff that has taken place between the office administrators. The two are pretty much available but are not talking to each other which, in other words, means that the updates are not flowing from one to the other. *We have to be partition tolerant!!!* As a training coordinator, you instruct them saying that none of you are taking any calls requesting for schedules or updates to schedules till you patch up. This implies that the system is partition tolerant but not available at that time.

In summary, one can at most decide to go with two of the three.

1. **Consistent:** The instructors or the training coordinator, once they have updated information with you, will always get the most updated information when they call subsequently.
2. **Availability:** The instructors or the training coordinators will always get the schedule if any or both of the office administrators have reported to work.
3. **Partition Tolerance:** Work will go on as usual even if there is communication loss between the office administrators owing to a spat or a tiff!

When to choose consistency over availability and vice-versa…

1. Choose availability over consistency when your business requirements allow some flexibility around when the data in the system synchronizes.
2. Choose consistency over availability when your business requirements demand atomic reads and writes.

Examples of databases that follow one of the possible three combinations:

1. Availability and Partition Tolerance (AP)
2. Consistency and Partition Tolerance (CP)
3. Consistency and Availability (CA)

Refer Figure 3.15 to get a glimpse of databases that adhere to two of the three characteristics of CAP theorem.

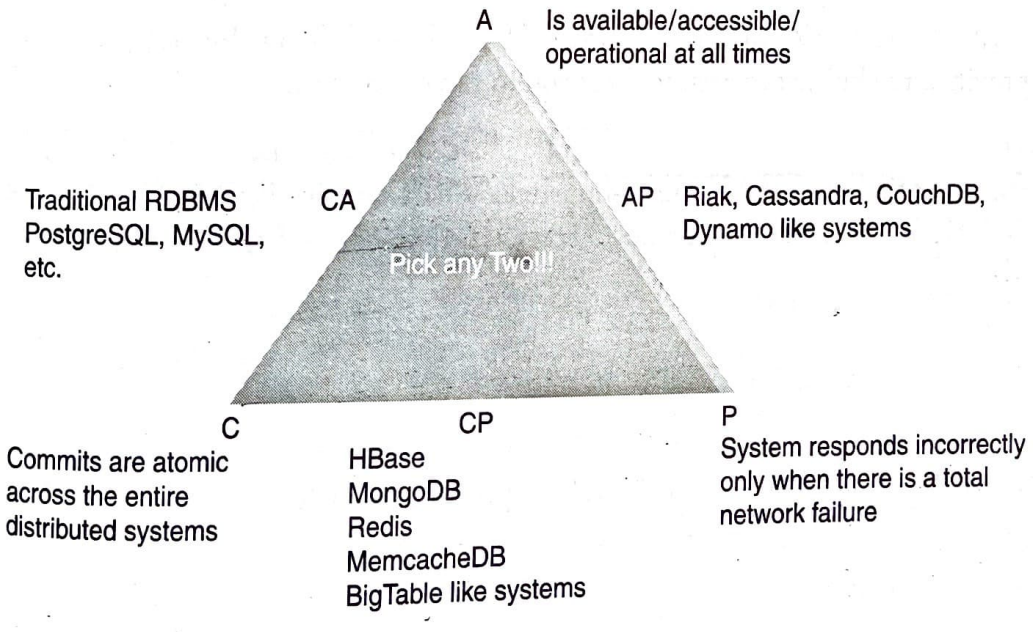


Figure 3.15 Databases and CAP.

3.13 BASICALLY AVAILABLE SOFT STATE EVENTUAL CONSISTENCY (BASE)

A few basic questions to start with:

1. ***Where is it used?***
   In distributed computing.
2. ***Why is it used?***
   To achieve high availability.
3. ***How is it achieved?***
   Assume a given data item. If no new updates are made to this given data item for a stipulated period of time, eventually all accesses to this data item will return the updated value. In other words, if no new updates are made to a given data item for a stipulated period of time, all updates that were made in the past and not applied to this given data item and the several replicas of it will percolate to this data item so that it stays as current/recent as is possible.
4. ***What is replica convergence?***
   A system that has achieved eventual consistency is said to have converged or achieved *replica convergence.*
5. ***Conflict resolution*: How is the conflict resolved?**
   (a) **Read repair:** If the read leads to discrepancy or inconsistency, a correction is initiated. It slows down the read operation.
   (b) **Write repair:** If the write leads to discrepancy or inconsistency, a correction is initiated. This will cause the write operation to slow down.
   (c) **Asynchronous repair:** Here, the correction is not part of a read or write operation.

3.14 FEW TOP ANALYTICS TOOLS

There is no dearth of analytical tools in the market. Please find below our list of few top analytics tools. We have also provided the links after each tool for you to explore more…

1. MS Excel
   https://support.office.microsoft.com/en-in/article/Whats-new-in-Excel-2013-1cbc42cd-bfaf-43d7-9031-5688ef1392fd?CorrelationId=1a2171cc-191f-47de-8a55-08a5f2e9c739&ui=en-US&rs=en-IN&ad=IN
2. SAS
   http://www.sas.com/en_us/home.html
3. IBM SPSS Modeler
   http://www-01.ibm.com/software/analytics/spss/products/modeler/
4. Statistica
   http://www.statsoft.com/

5. Salford systems (World Programming Systems)
   http://www.salford-systems.com/
6. WPS
   http://www.teamwpc.co.uk/products/wps

3.14.1 Open Source Analytics Tools

Let us look at a couple of open source analytics tools. We have also provided the links after each tool for you to explore more…

1. R analytics
   http://www.revolutionanalytics.com/
2. Weka
   http://www.cs.waikato.ac.nz/ml/weka/

REMIND ME

- 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.
- Big data analytics is a about a tight handshake between three communities: IT, business users, and data scientists.
- *Data science* is the science of extracting knowledge from data.
- The CAP theorem is also called the Brewer's Theorem. It states that in a distributed computing environment (a collection of interconnected nodes that share data), it is impossible to provide the following guarantees. At best you can have two of the following three – one must be sacrificed.
  - Consistency
  - Availability
  - Partition tolerance

CONNECT ME (INTERNET RESOURCES)

- http://en.wikipedia.org/wiki/Data_science
- http://simplystatistics.org/2013/12/12/the-key-word-in-data-science-is-not-data-it-is-science/
- http://www.oralytics.com/2012/06/data-science-is-multidisciplinary.html
- http://spotfire.tibco.com/blog/?p=4240
- http://reports.informationweek.com/abstract/106/1255/Financial/tech-center-taking-advantage-of-in-memory-analytics.html
- http://www.informationweek.com/software/information-management/oracle-analytics-package-expands-in-database-processing-options/d/d-id/1102712?

TEST ME

A. Fill Me

1. The _____ technology helps query data that resides in a computer's random access memory (RAM) rather than data stored on physical disks.
2. Eventual consistency is a consistency model used in distributed computing to achieve high _____.
3. A coordinated processing of a program by multiple processors, each working on different parts of the program and using its own operating system and memory is called _____.
4. A collection of independent computers that appear to its users as a single coherent system is _____.

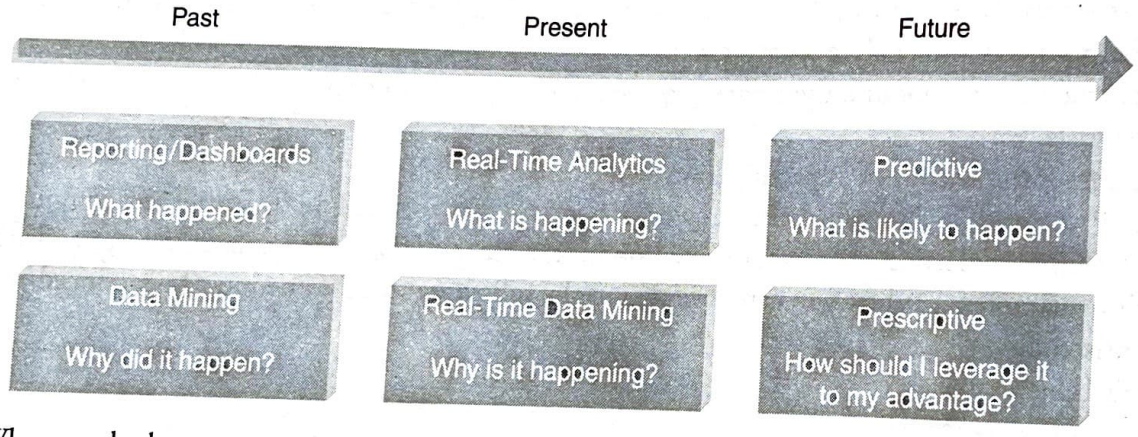
Answers:

1. In-memory analytics
2. Availability
3. Massively parallel processing
4. Distributed systems

B. Answer Me

1. What are the various types of analytics?

Answer:



2. What are the key questions to be answered by all organizations stepping into analytics?

Answer: The key questions for any organization stepping into analytics are:

- Should you be storing all of your big data? If "Yes", where are you going to store it? If "No", how will you know what to store and what to discard?
- How will you sieve through your massive data to filter out the relevant from the irrelevant?
- How long will you store this data?
- How will you accommodate the peaks (variability in terms of data influx) in your data?
- How will you analyze? Will you analyze all the data that is stored or analyze a sample?
- What will you do with the insights generated from this analysis?

3. What can one expect from analytics 3.0?

Answer:

- In-memory analytics.
- In-database processing.
- Leveraging analytics to improve operational, tactical, and strategic decision making.

- Coupling the in-memory analytics and in-database processing with agile analytical methods and machine learning techniques.
- Appropriate tools to effectively support decision-making at the front lines, such as mobile and self-serve analytical applications.

4. Which industries will be affected most by analytics 3.0? Who will benefit the most?

Answer: Almost all the firms in all the industries and not just online firms will be affected by analytics 3.0. A lot of analytics have already been done in the Transport, Retail, and Banking sector. Telecom, entertainment, and health sectors have a bit of catching up to do.

5. What is predictive and prescriptive analytics?

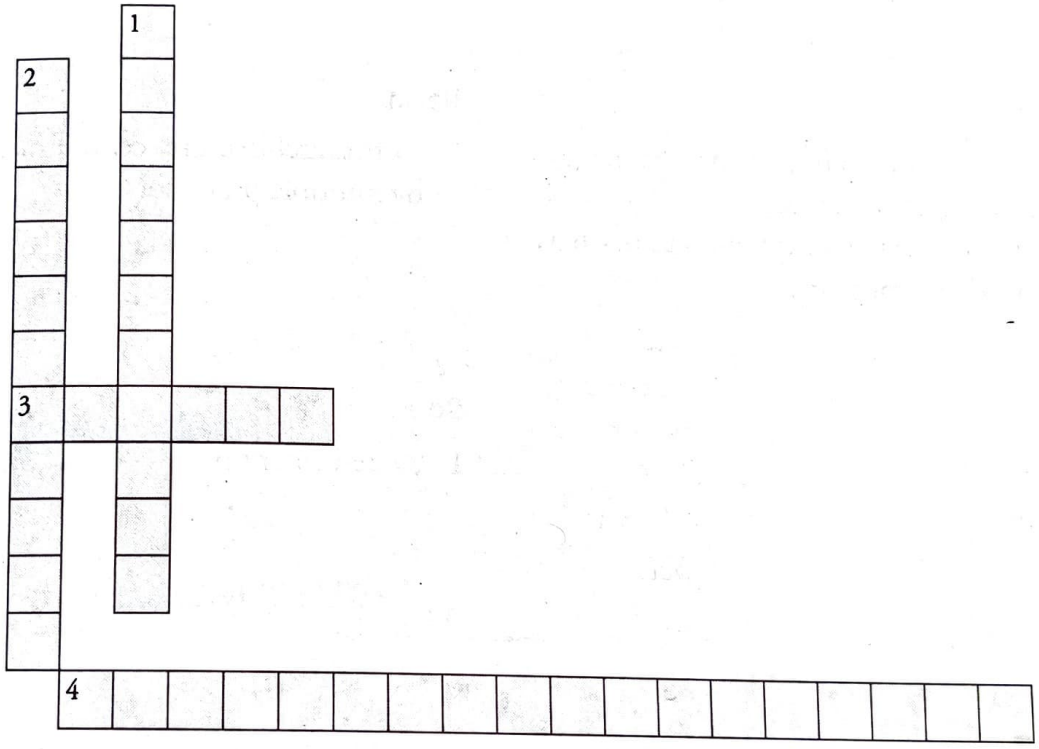
Answer:

Predictive analytics helps you answer the questions: "What will happen?" and "Why will it happen"?

Prescriptive analytics goes beyond "What will happen?" "Why will it happen?" and "When will it happen?" to answer "What should be the action taken to take advantage of what will happen"?

C. Crossword

1. Puzzle on CAP Theorem



Across

3. CAP theorem is also called as _____ theorem.
4. System will continue to function even when network partition occurs.

Down

1. Every read fetches the most recent write.
2. A non-failing node will return a reasonable response within a reasonable amount of time.

Solution:

Across

3. Brewer
4. Partition Tolerant

Down

1. Consistency
2. Availability