SESSION 2016-2017

B.TECH (CSE) YEAR: IV SEMESTER: VIII

BIG DATA AND ANALYTICS (CSE448)

MODULE 1 (L3)

Presented By

Dilip Kumar Sharma , Rahul Pradhan, Vivek Kumar, Yogesh Gupta

Dept of Computer Engineering & Applications

GLA University India

Objective vs. Outcomes

Learning Objectives	Learning Outcomes
Big Data Analytics	
1. What is big data analytics and what it isn’t?	a) To understand the significance of big data analytics.
2. Why is big data analytics important?	b) To understand the role of data scientist.
3. What is data Science?	c) To understand the various terminologies used in the big data environment.
4. Getting familiar with the terminologies used in the big data environment.	

Agenda

- What is Big Data Analytics?
- What Big Data Analytics isn’t?
- Classification of Analytics
- Why is Big Data Analytics Important?
- Data Science
- Data Scientist … Your New Best Friend!!!
- Terminologies Used in Big Data Environment
  - In Memory Analytics
  - In Database Processing
  - Massively Parallel Processing
  - Difference between Parallel versus Distributed Systems
  - Shared Nothing Architecture
  - Consistency, Availability, Partition Tolerance (CAP): Theorem Explained
- Few Top Analytics Tools

Scenarios

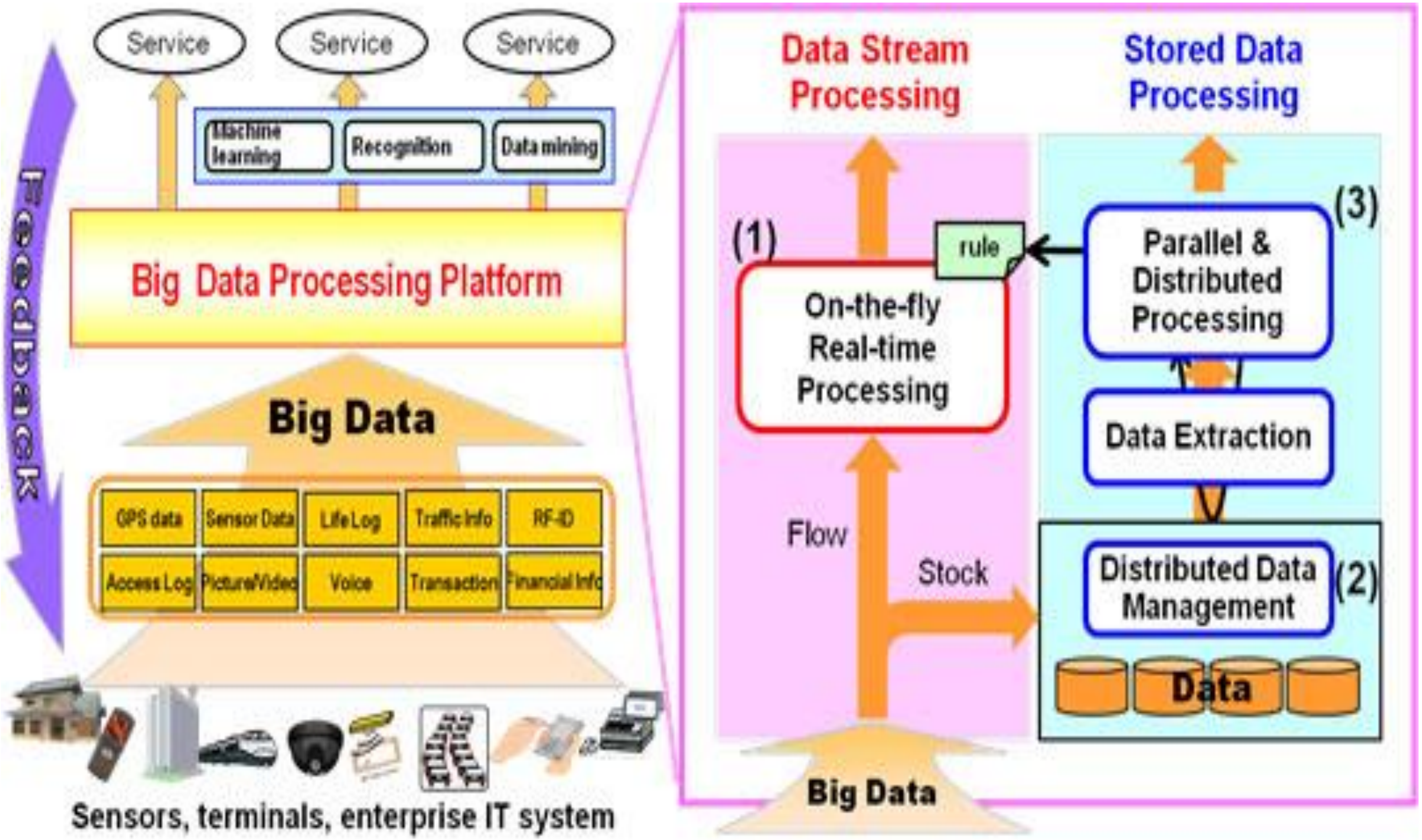
- **Scenario 1:** You have heard a Lot from your friends about the deals on offer on the Amazon site. You make a purchase on their site. There is something that does not escape your attention. Amazon has made a few suggestions (of books on similar topics or books by the same author) to you to help with 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?

Scenarios cont…

- **Scenario 2:** You are the owner of a trucks transport company. You get a call to help with a cargo delivery. They are ready to pay dou-ble the charge. You do not want to miss this oppor-tunity. 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.

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.


Service
Service
Service
Feedback
Machine learning
Recognition
Data mining
Big Data Processing Platform
Big Data
GPS data
Sensor Data
Life Log
Traffic Info
RF-ID
Access Log
Picture/Video
Voice
Transaction
Financial Info
Sensors, terminals, enterprise IT system
Data Stream Processing
Stored Data Processing
(1)
On-the-fly Real-time Processing
rule
(3)
Parallel & Distributed Processing
Data Extraction
Flow
Stock
Distributed Data Management
(2)
Data
Big Data

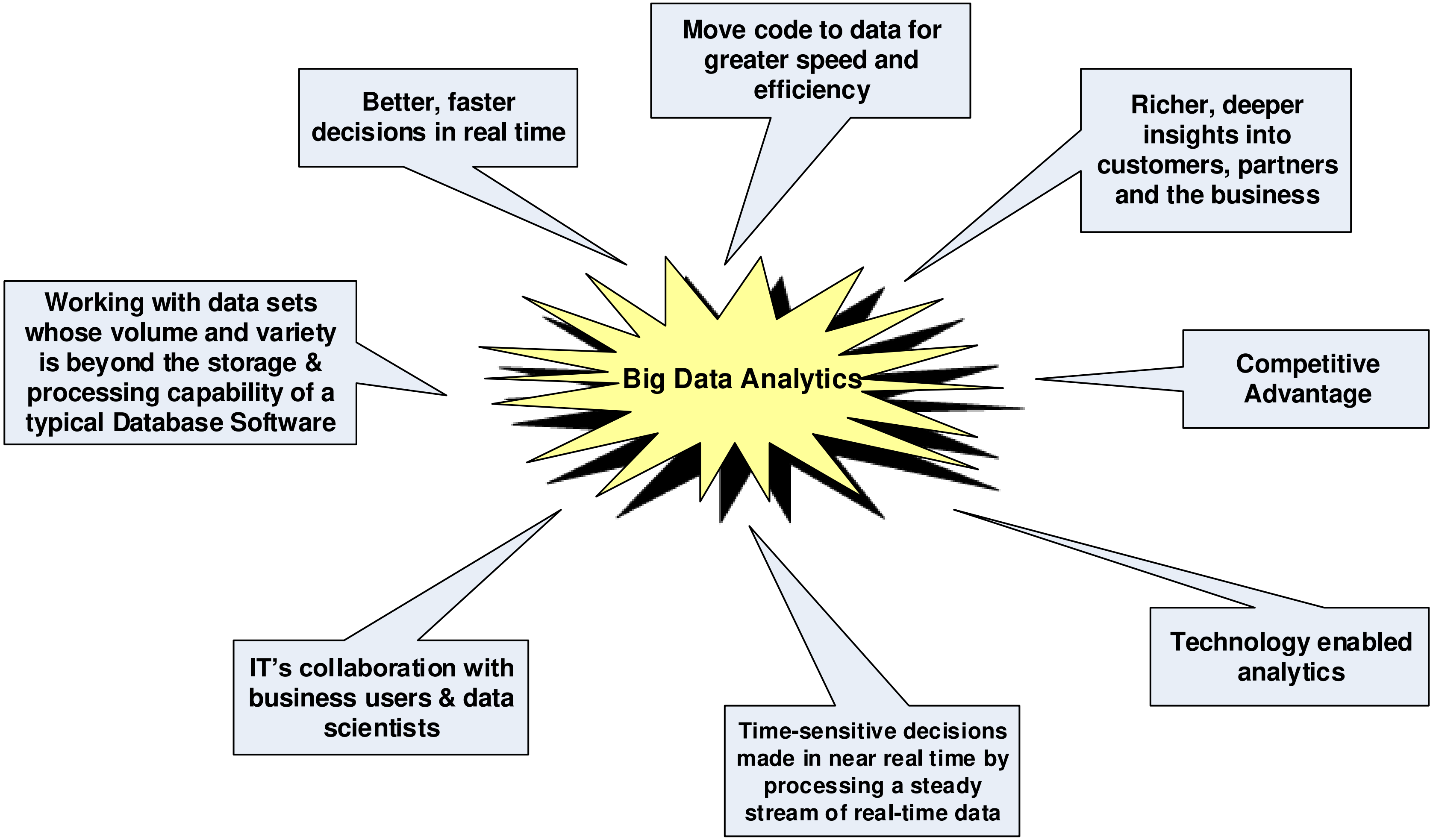
Big Data Analytics

- **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.
- About gaining a meaningful, deeper, and richer insight into your business to steer it in the right direction, understanding the customers demographics to cross-sell and up-sell to them.

Big Data Analytics cont…

- A tight handshake between three communities: IT, business users, and data scientists.
- Working with datasets whose volume and variety exceed the current storage and processing capabilities and infrastructure of your enterprise.
- 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 is Big Data Analytics?



Big Data Analytics

Better, faster decisions in real time

Move code to data for greater speed and efficiency

Richer, deeper insights into customers, partners and the business

Working with data sets whose volume and variety is beyond the storage & processing capability of a typical Database Software

Competitive Advantage

IT's collaboration with business users & data scientists

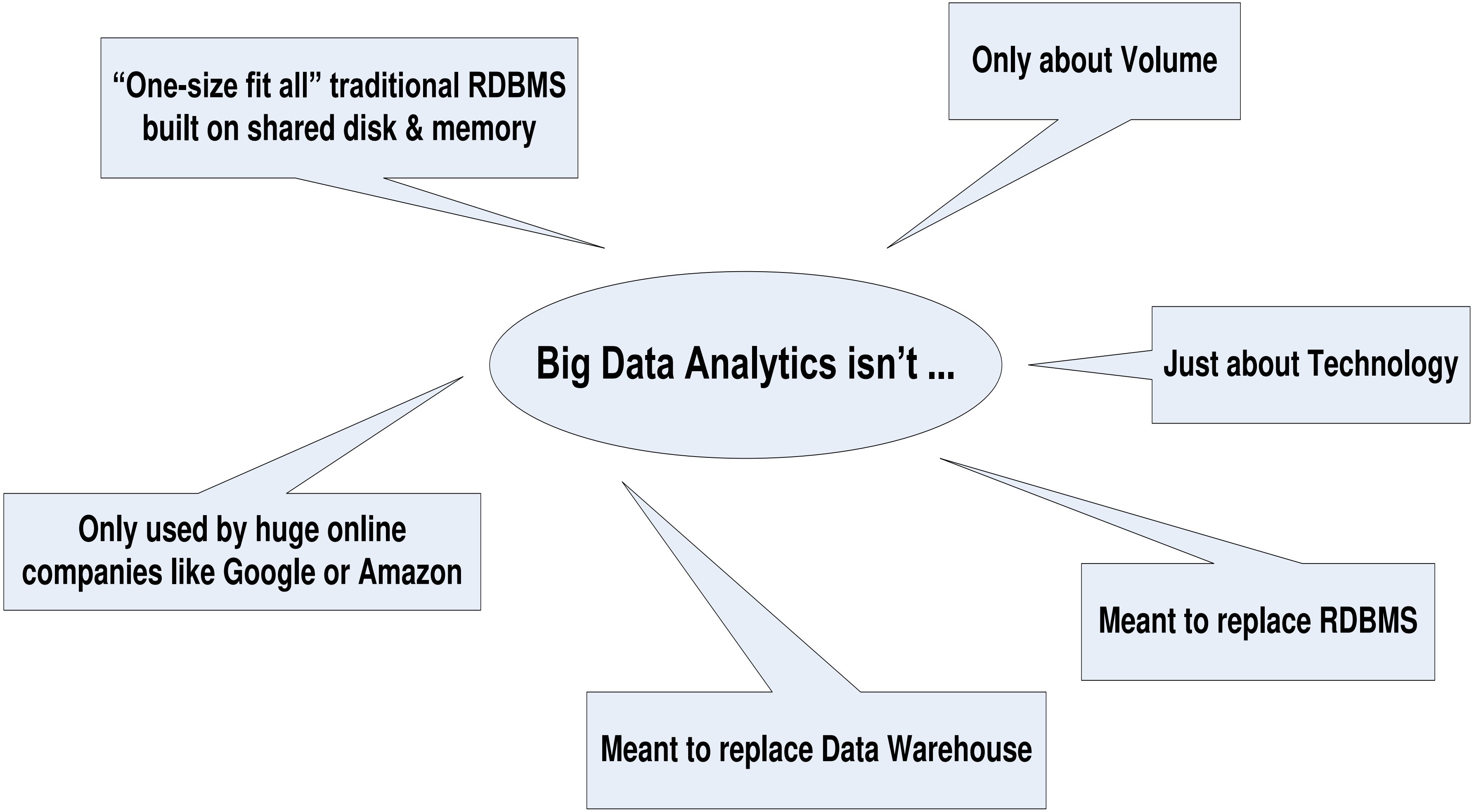
Time-sensitive decisions made in near real time by processing a steady stream of real-time data

Technology enabled analytics

What Big Data Isn’t?

- □ Big data isn’t just about technology. It is about understanding what the data is saying to us.
- □ It is about patterns and trends waiting to be unveiled.
- □ Meant to replace RDBMS ?
- □ Meant to replace data warehouse ?
- □ And of course, big data analytics is not here to replace our now very robust and powerful RDBMS or Data Warehouse. It is here to coexist with them.

What Big Data Analytics isn't?



Big Data Analytics isn't ...

"One-size fit all" traditional RDBMS built on shared disk & memory

Only about Volume

Just about Technology

Meant to replace RDBMS

Meant to replace Data Warehouse

Only used by huge online companies like Google or Amazon

Why this industry buzz?

- Let us put it down to three foremost reasons:
  - Data is growing at a 40% compound annual rate, reaching nearly 45 ZB by 2020.
  - The volume of business data worldwide is expected to double every 1.2 years.
  - 500 million “tweets” are posted by Twitter users every day.
  - 2.7 billion “Likes” and comments are posted by Facebook users in a day.
  - 90% of the world’s data created in the past 2 years.

Why this industry buzz? Cont…

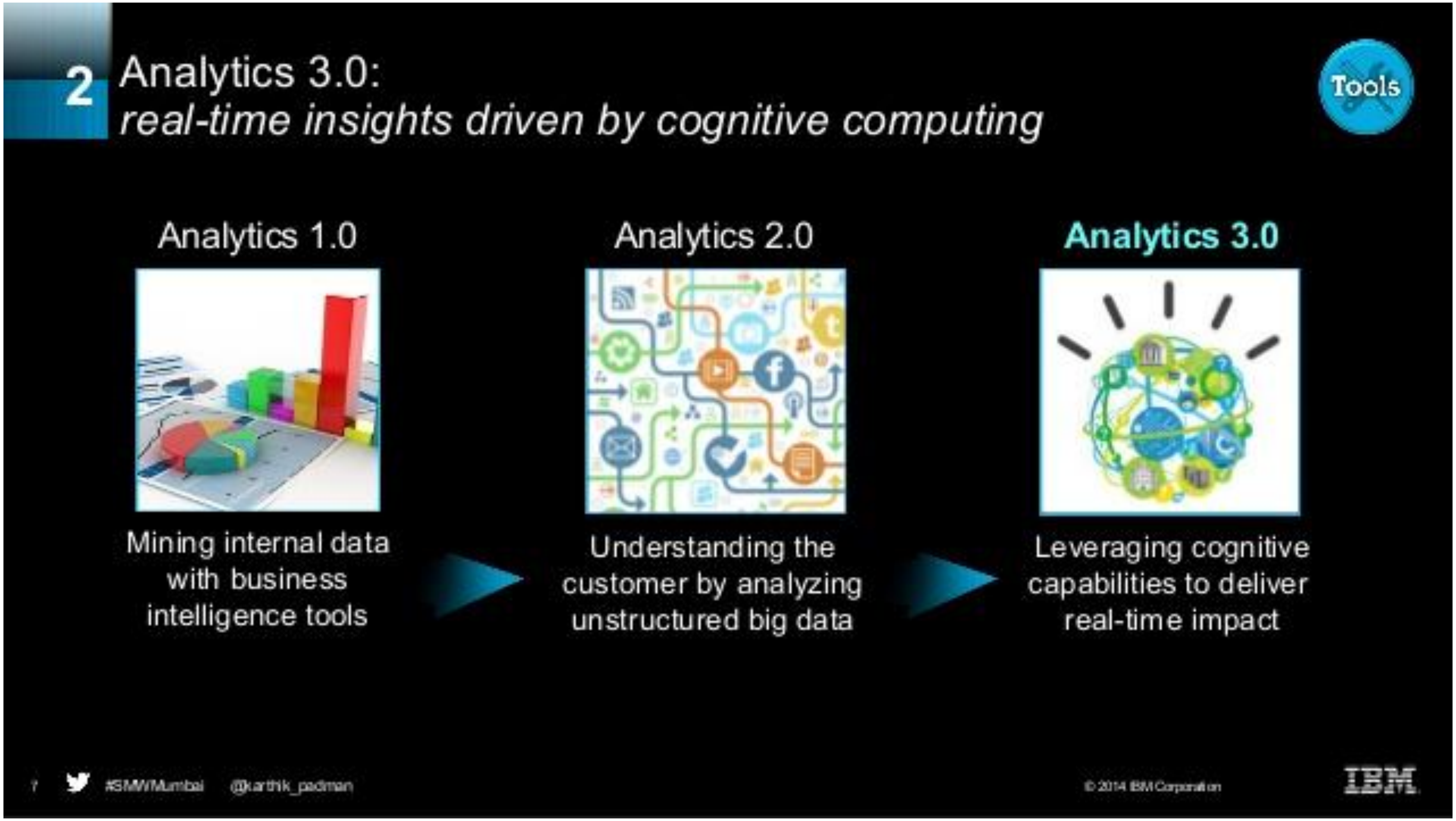
- Source:
  - http://www.intel.com/content/www/us/en/communications/internet-minute-infographic.html
  - http://www.ibm.com/software/data/bigdata/what-is-big-data.html
- Cost per gigabyte of storage has hugely dropped.
- There are an overwhelming number of user-friendly analytics tools available in the market today.

Classification of Analytics

- There are basically two schools of thought:
  - Those that classify analytics into basic, operationalized, advanced, and monetized.
  - Those that classify analytics into analytics 1.0, analytics 2.0, and analytics 3.0.

First School of Thought

- □ **Basic analytics:** This primarily is slicing and dicing of data to help with basic business insights.
- □ **Operationalized analytics:** It is operationalized analytics if it gets woven into the enterprise's business processes.
- □ **Advanced analytics:** This largely is about forecasting for the future by way of predictive and prescriptive modeling. .
- □ **Monetized analytics:** This is analytics in use to derive direct business revenue.



2 Analytics 3.0:
real-time insights driven by cognitive computing

Analytics 1.0

Mining internal data with business intelligence tools

Analytics 2.0

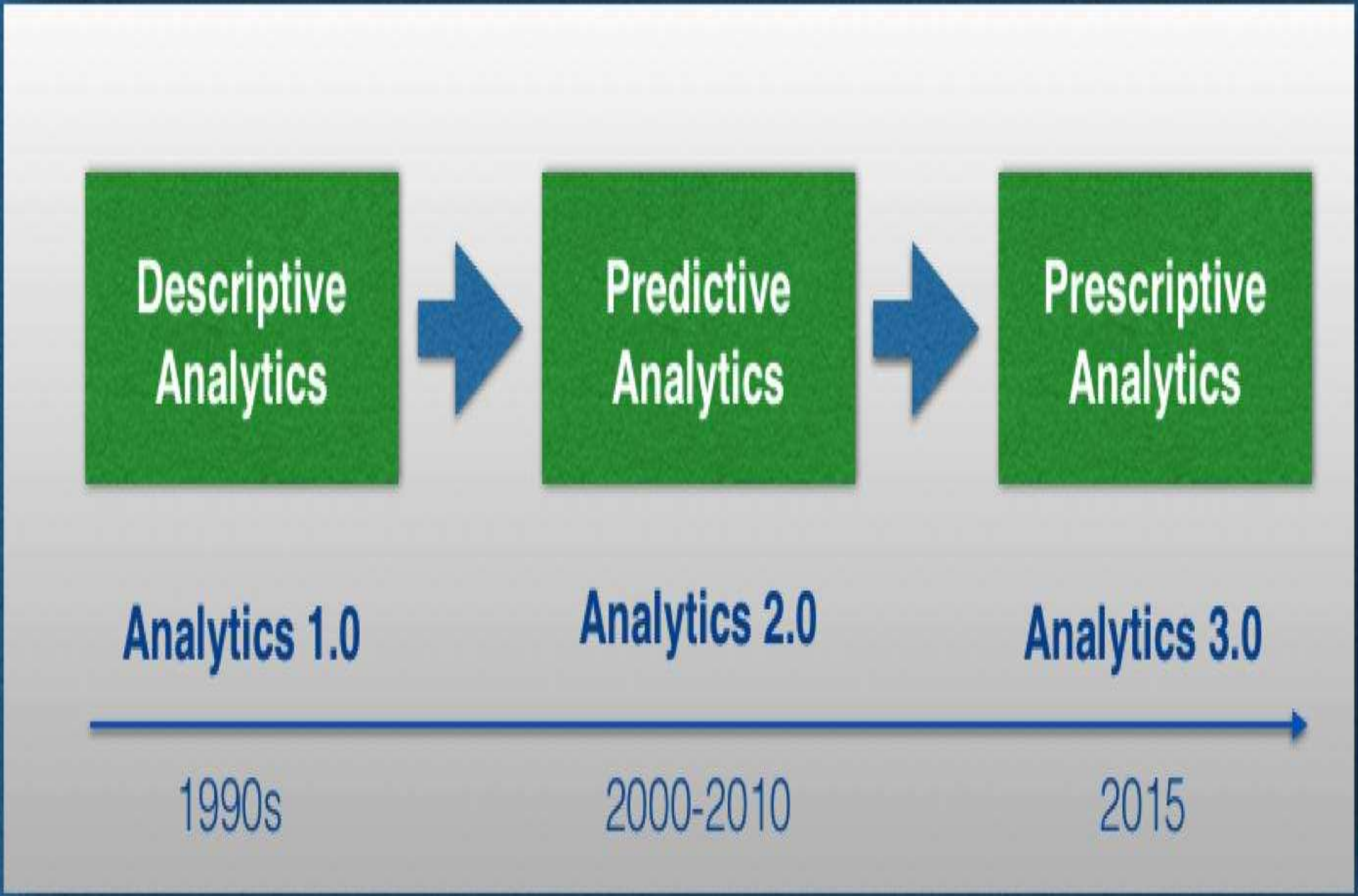
Understanding the customer by analyzing unstructured big data

Analytics 3.0

Leveraging cognitive capabilities to deliver real-time impact

Second School of Thought

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)
Relational databases	Database appliances, Hadoop clusters, SQL to Hadoop environments, etc.	In memory analytics, in database processing, agile analytical methods, machine learning techniques, etc.



Descriptive Analytics

Predictive Analytics

Prescriptive Analytics

Analytics 1.0

Analytics 2.0

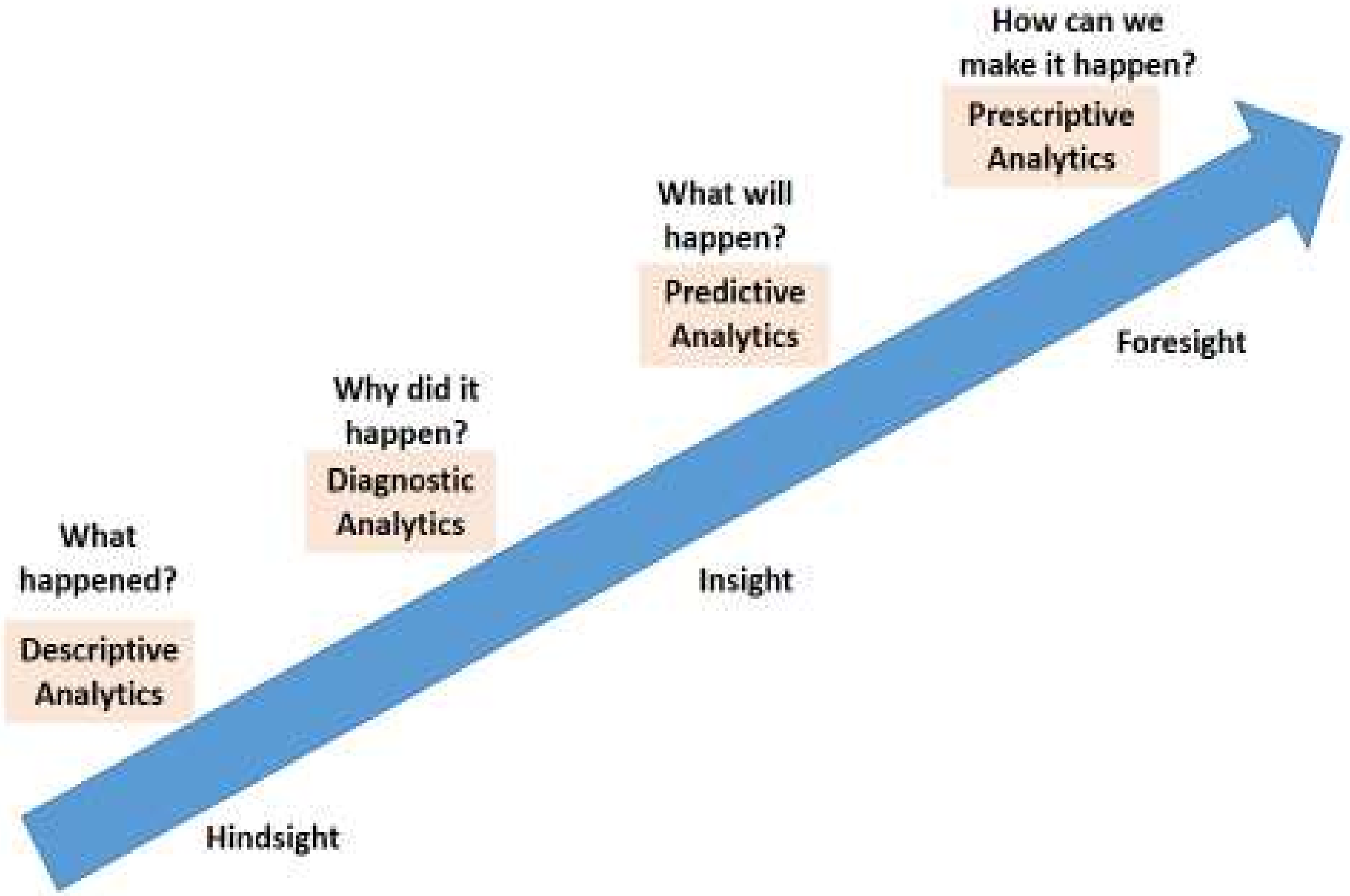
Analytics 3.0

1990s

2000-2010

2015

Analytics 1.0, 2.0 and 3.0



What happened?

Descriptive Analytics

Why did it happen?

Diagnostic Analytics

What will happen?

Predictive Analytics

How can we make it happen?

Prescriptive Analytics

TOP CHALLENGES : BIG DATA

- **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?

TOP CHALLENGES cont...

- **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 thai cannot be ignored given that big data carries credit card information, personal information, and other sensitive data.

TOP CHALLENGES cont...

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

TOP CHALLENGES cont...

- **Consistency:** Should one opt for consistency or eventual consistency?
- **Partition tolerant:** How to build partition tolerant systems that can take care of both hardware and software failures?
- **Data quality:** How to maintain data quality - data accuracy, completeness, timeliness, etc.? Do we have appropriate metadata in place?

WHY ANALYTICS IMPORTANT?

- The various approaches to analysis of data and what it leads to.
  - **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.

ANALYTICS IMPORTANCE cont...

- **Reactive - Big Data Analytics:** Here the analysis is done on huge datasets but the approach is sdl reactive as it is still based on static data.

ANALYTICS IMPORTANCE cont…

- **Proactive Analytics:** This is to support futuristic decision making by the use of data mining, pre-dictive 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.

ANALYTICS IMPORTANCE cont…

- **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.

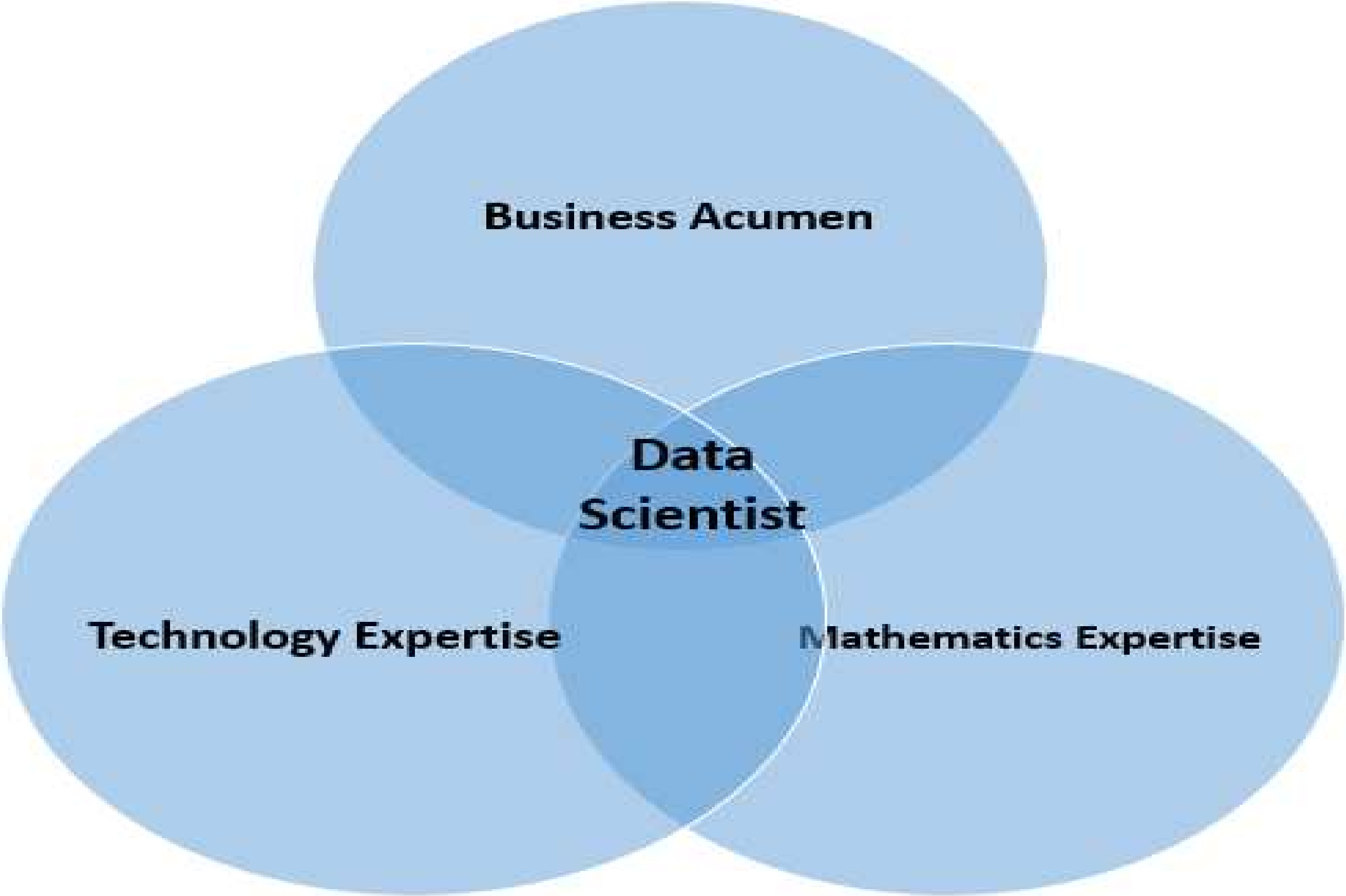
CHALLENGES POSED BY BIG DATA

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

Data Science

- *Data science* is the science of extracting knowledge from data. It employs techniques and theories drawn from many fields from the broad areas of mathematics, statistics, information technology including machine ng. 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 frauds, terrorist network and activities, global economic impacts, sensor logs, social media analytics, and so many areas.
- Data science is multi-disciplinary.

Data Scientist



Business Acumen Skills

- A data scientist should have business acumen skills to counter the pressure of business:
  - Understanding of domain
  - Business strategy
  - Problem solving
  - Communication
  - Presentation
  - Inquisitiveness

Technology Expertise Skills

- A data scientist should be technology expert to convert the business into business logic:
  - Good database knowledge such as RDBMS.
  - Good NoSQL database knowledge such as MongoDB, Cassandra, HBase, etc.
  - Programming languages such as Java, Python, etc.
  - Open-source tools such as Hadoop, R.
  - Datawarehousing, Datamining.
  - Visualization such as Tableau, Flare, Google visualization APIs, etc.

Mathematics Expertise Skills

- A data scientist should be mathematics expert to formulize and analyze data:
  - Mathematics.
  - Statistics.
  - Artificial Intelligence (AI).
  - Machine learning.
  - Pattern recognition.
  - Natural Language Processing.

Data Science Process

- □ Collecting raw data from multiple data sources.
- □ Processing the data.
- □ Integrating the data and preparing clean datasets.
- □ Engaging in explorative data analysis using model and algorithms.
- □ Preparing presentations using data visualizations (commonly called Infographics, or BizAnalytics, etc.)
- □ Communicating the findings to all stakeholders.
- □ Making faster and better decisions.

Responsibilities of Data Scientist

- **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.

Responsibilities cont…

- **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.

Responsibilities cont…

- **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.

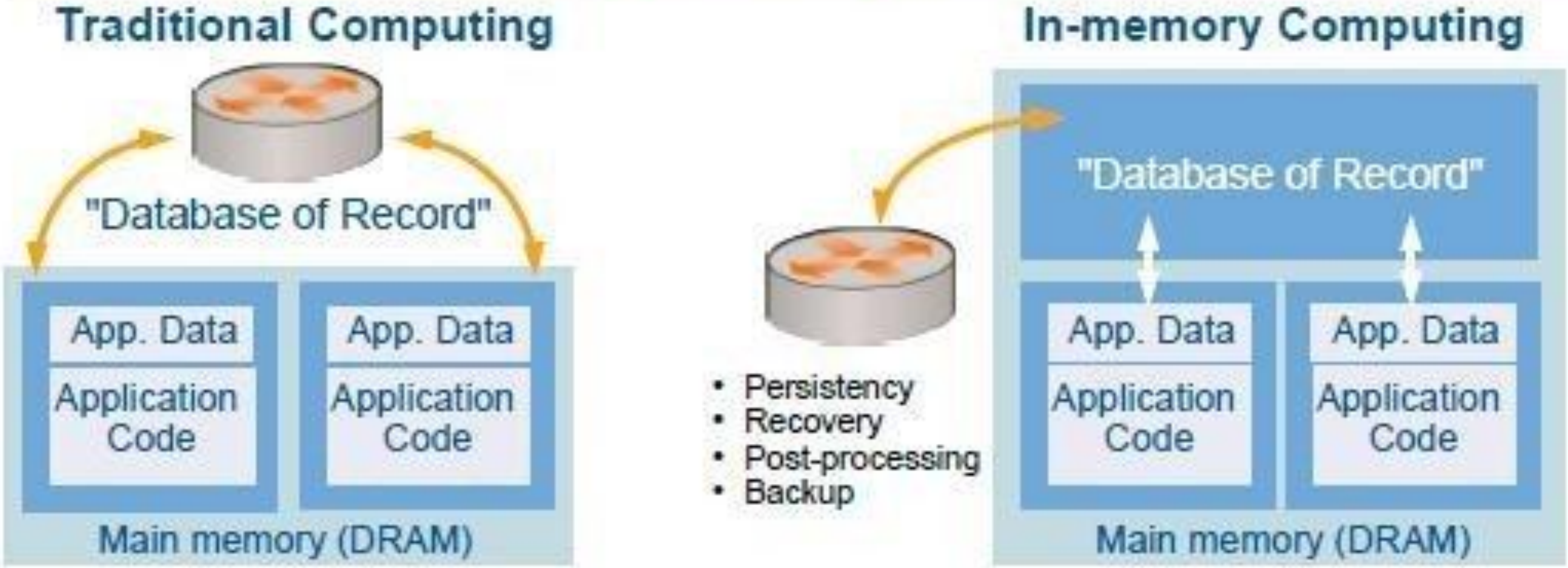
Terminologies Used in Big data Environments

- In-Memory Analytics
- In-Database Processing
- Massively Parallel Processing
- Parallel System
- Distributed System
- Shared Nothing Architecture

In-Memory Analytics

- Data access from non-volatile storage such as hard disk is a slow process. 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.
- This problem has been addressed using in-memory analytics. Here all the relevant data is stored in RAM.
- The advantage is faster access, rapid deployment, better insights, and minimal IT involvement.

What Is In-memory Computing?



- Persistency
- Recovery
- Post-processing
- Backup

Why Now?

- 64-bit processors can address **up to 16 exabytes of data**
- DRAM production costs **drop by 32% every 12 months**
- 1GB of NAND flash memory **average price is 56$ cents***
- Commodity hardware provide **multi terabyte of DRAM**
- In-memory-enabling **software is available and proven**
- IMC software is often **embedded in products/services**

* Per Gartner's "Weekly Memory Pricing Index, 21 December 2012," G00247628

Gartner.

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 OLTP systems after cleaning up (de-duplication, scrubbing, etc.) through the process of ETL is stored in the Enterprise Data Warehouse (EDW) or data marts.
- With in-database processing, the database program itself can run the computations eliminating the need for export and thereby saving on time.

Symmetric Multiprocessor System

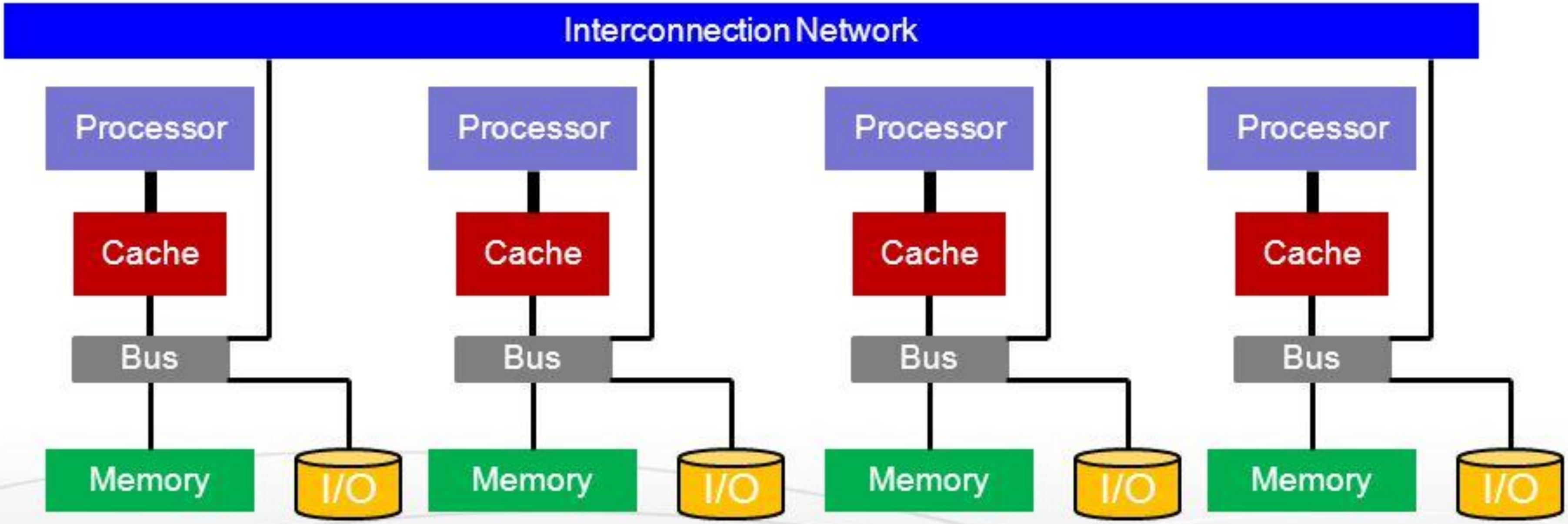
- In SMP, there is a single common main memory that is shared by two or more identical processors. The processors have full access to all I/O devices and are controlled by a single operating system instance.
- SMP are tightly coupled multiprocessor systems. Each processor has its own high-speed memory, called cache memory and are connected using a system bus.

Massively 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.

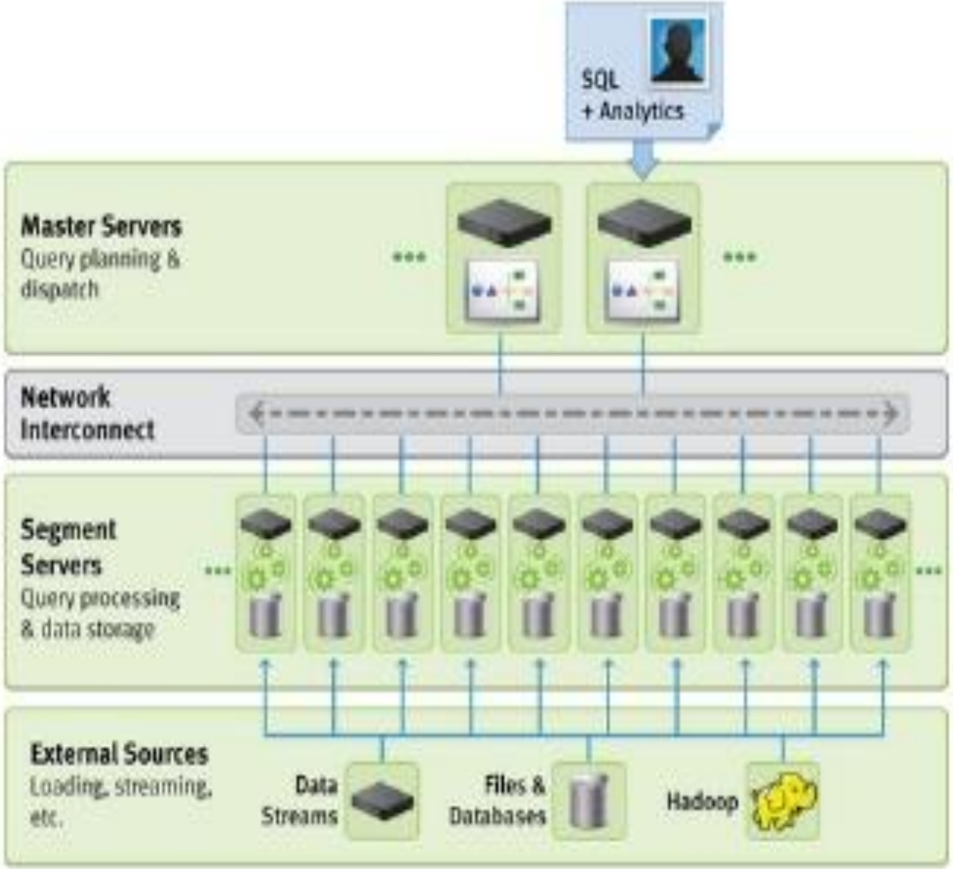
Massively Parallel Processors

- Massively Parallel Processors (MPP) architecture consists of nodes with each having its own processor, memory and I/O subsystem

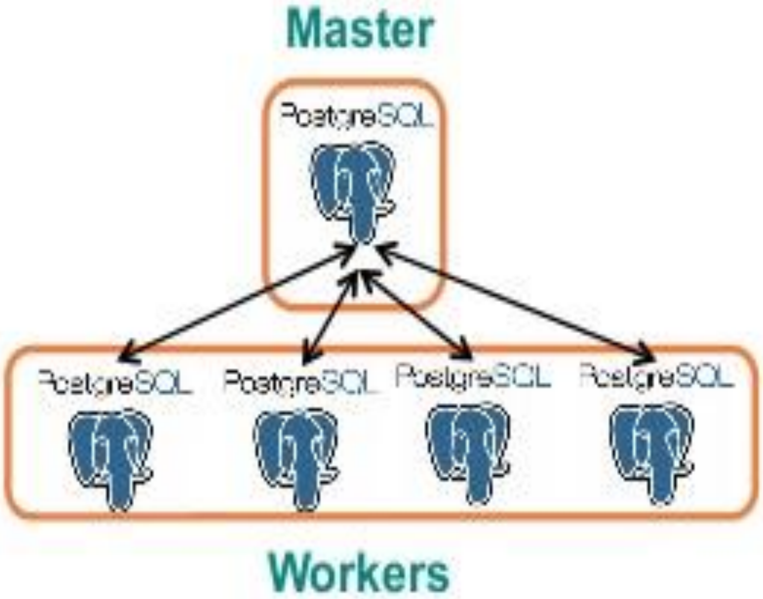


- An independent OS runs at each node

MPP Architectural Overview



Think of it as multiple PostGreSQL servers



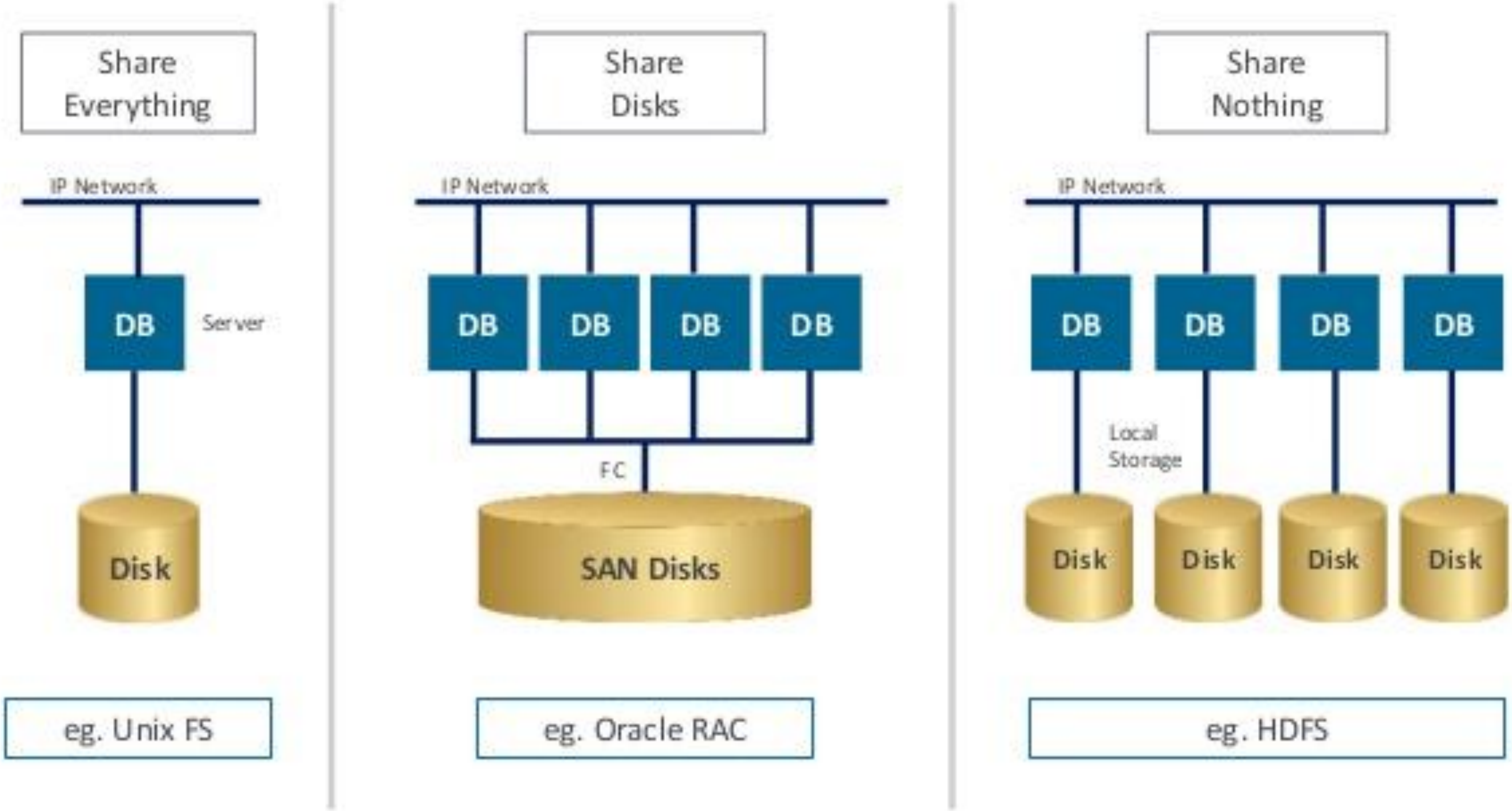
Parallel and Distributed System

- A parallel database system is a tightly coupled system. The processor, co-operate for query processing. The user is unaware of the parallelism.
- Distributed database systems are known to be loosely coupled and are composed by individual machines. Each of the machines can run their individual application and serve their own respective users. The data is usually distributed across several machines.

Shared Nothing Architecture

- In shared nothing architecture, neither memory nor disk is shared among multiple processors.
- Advantages:
  - Fault Isolation
  - Scalability

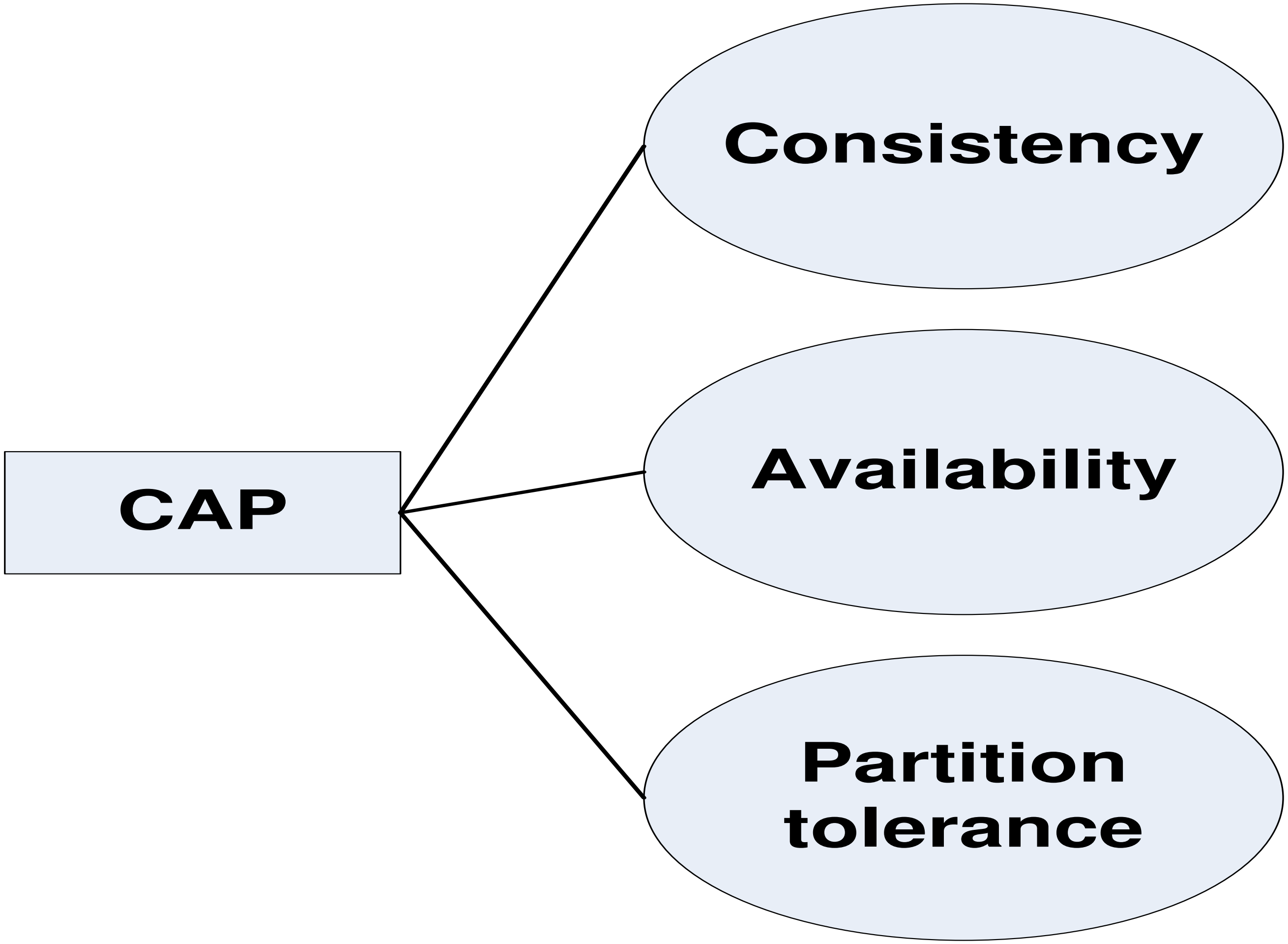
SHARE NOTHING ARCHITECTURE

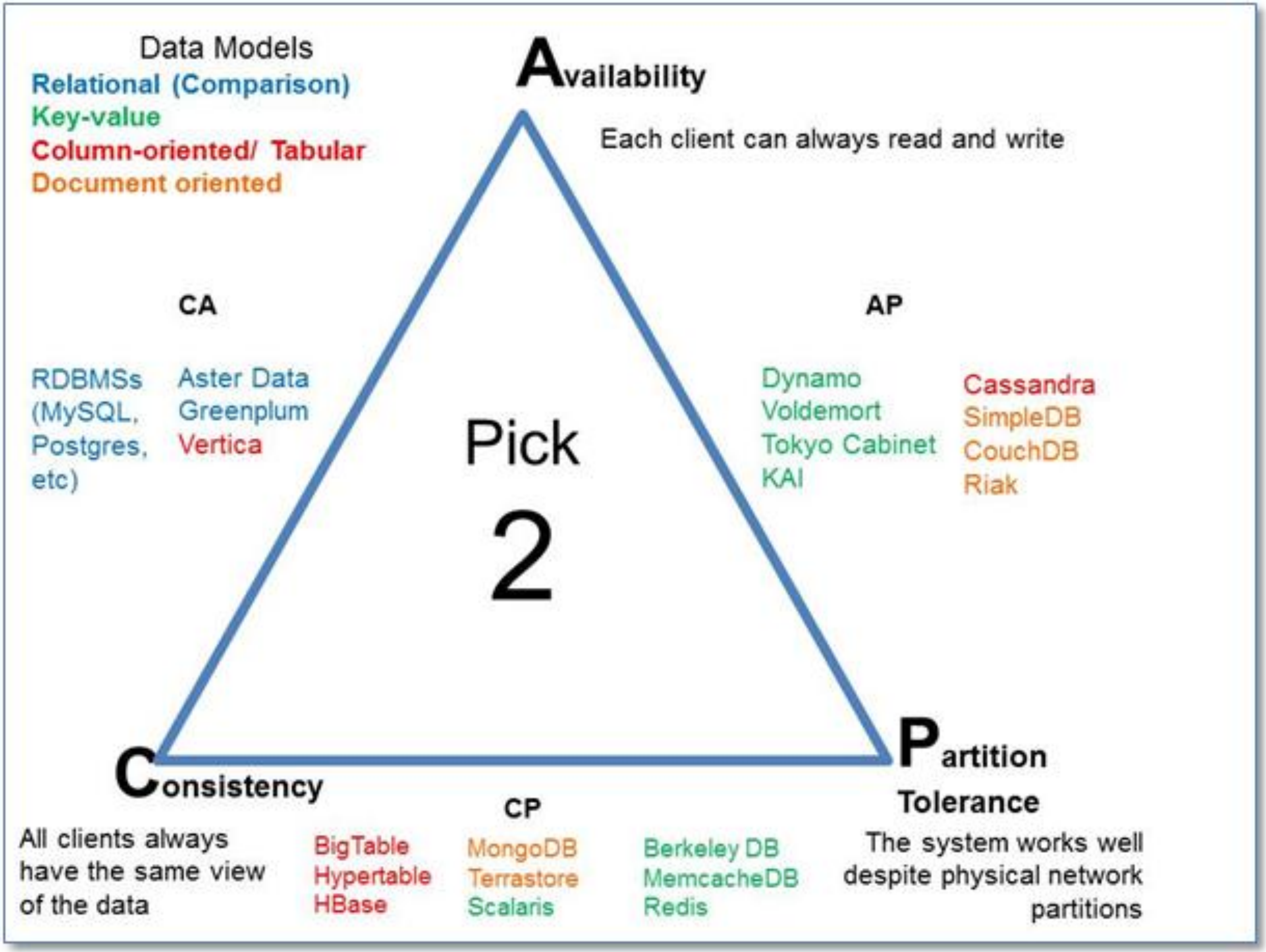


CAP Theorem

- The CAP Theorem is also called *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 gaurantees:
  - Consistency
  - Availability
  - Partition tolerance

Brewer’s CAP





Data Models

Relational (Comparison)

Key-value

Column-oriented/ Tabular

Document oriented

Availability

Each client can always read and write

CA

RDBMSs (MySQL, Postgres, etc)

Aster Data
Greenplum
Vertica

AP

Dynamo
Voldemort
Tokyo Cabinet
KAI

Cassandra
SimpleDB
CouchDB
Riak

Pick 2

Consistency

All clients always have the same view of the data

CP

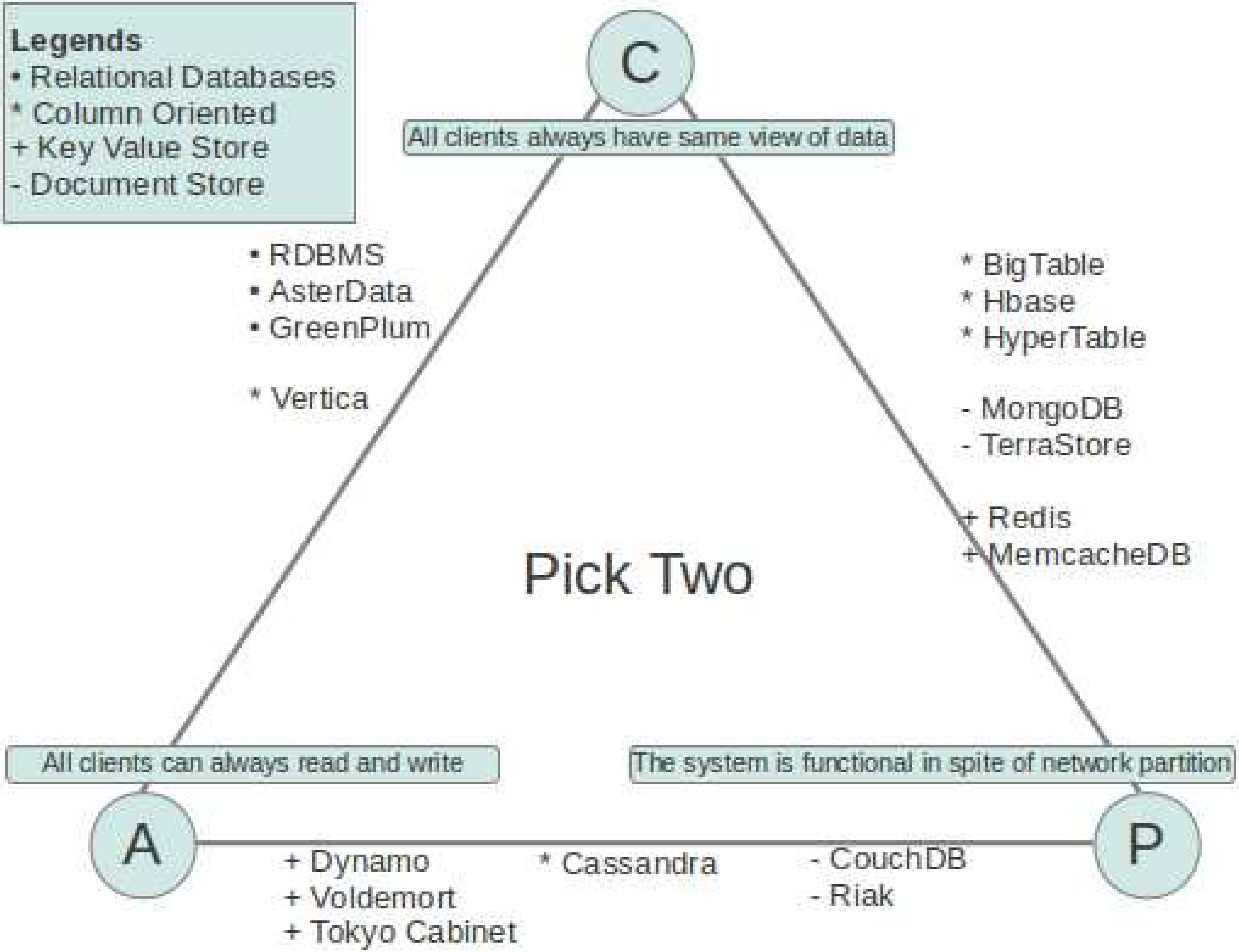
BigTable
Hypertable
HBase

MongoDB
Terrastore
Scalaris

Berkeley DB
MemcacheDB
Redis

Partition Tolerance

The system works well despite physical network partitions



Legends
• Relational Databases
* Column Oriented
+ Key Value Store
- Document Store

C

All clients always have same view of data

• RDBMS
• AsterData
• GreenPlum

* Vertica

* BigTable
* Hbase
* HyperTable

- MongoDB
- TerraStore

+ Redis
+ MemcacheDB

Pick Two

All clients can always read and write

The system is functional in spite of network partition

A

P

+ Dynamo
+ Voldemort
+ Tokyo Cabinet

* Cassandra

- CouchDB
- Riak

Few Top Analytical Tools

- 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

- SAS

  http://www.sas.com/en_us/home.html

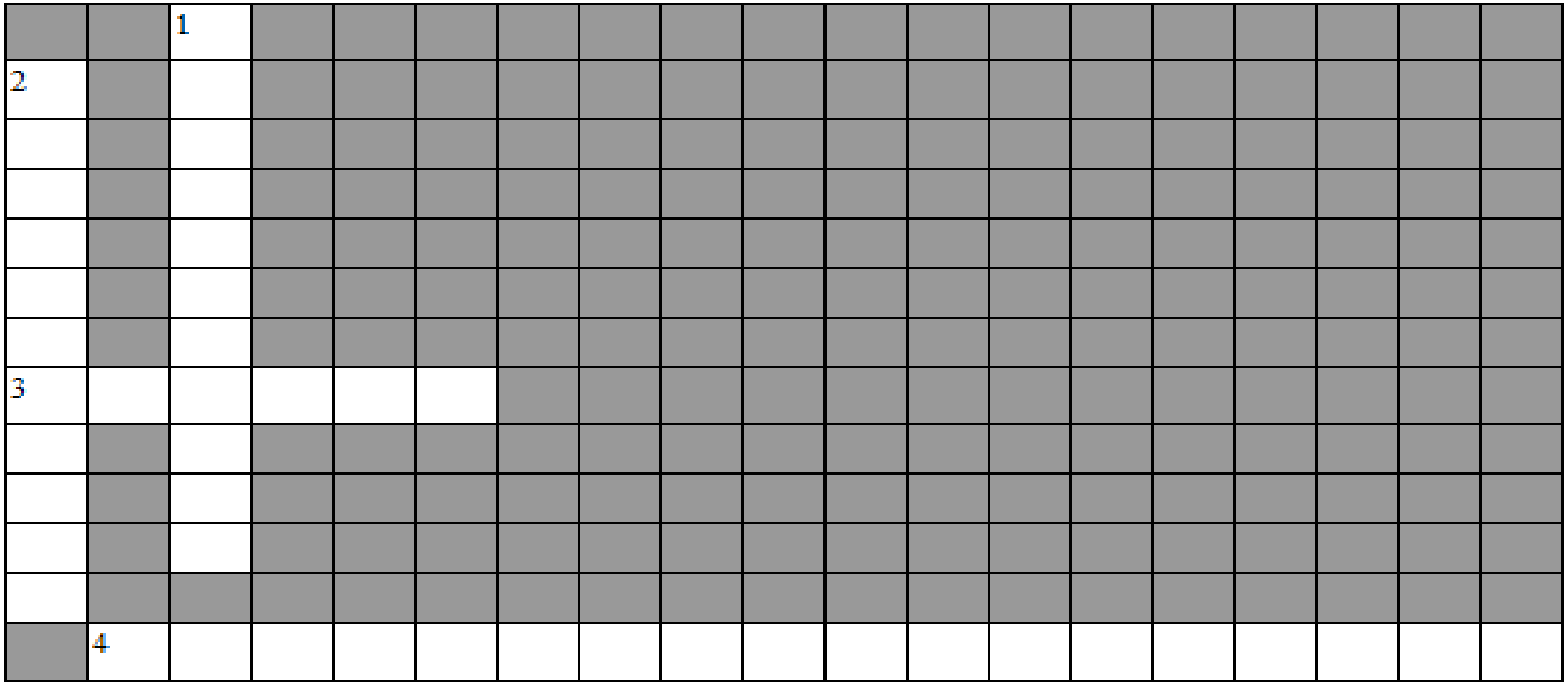
- IBM SPSS Modeler

  http://www-01.ibm.com/software/analytics/spss/products/modeler/

Answer a few quick questions …

- The _______ technology helps query data that resides in a computer's RAM rather than data stored in physical disks.
- 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 _________.
- A collection of independent computers that appear to its users as a single coherent system is _________.

Crossword 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

Question’s Answer ??

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

- What is predictive and prescriptive analytics?