FUNDAMENTALS OF BIG DATA ANALYTICS

UNIT-1

Types of Digital Data: Classification of Digital Data.
Introduction to Big Data: Characteristic of Data, Evolution of Big Data, Definition of Big Data, Challenges with Big Data, What is Big Data?.
Big Data Analytics: Where do we Begin?, What is Big Data Analytics?, What Big Data Analytics isn't?, Classification of Analytics, Terminologies Used in Big Data Environments.
The Big Data Technology Landscape: NoSQL

Q) What are the characteristics of data?

- Data is a collection of details in the form of either figures or texts or symbols, or descriptions etc.
- Data contains raw figures and facts. Information unlike data provides insights analyzed through the data collected.

Data has 3 characteristics:

1. Composition: The composition of data deals with the structure of data, i.e; the sources of data, the granularity, the types and nature of data as to whether it is static or real time streaming.
2. Condition: The condition of data deals with the state of data, i.e; "Can one use this data as is for analysis?" or "Does it require cleaning for further enhancement and enrichment?" data?"
3. Context: The context of data deals with "Where has this data been generated?". "Why was this data generated?", "How sensitive is this data?", "What are the events associated with this".

Q) What is digital data? Explain different types of digital data.

The data that is stored using specific machine language systems which can be interpreted by various technologies is called **digital data**.
Eg. Audio, video or text information

Digital Data is classified into three types:

1. **Structured Data:-**
   This is the data which is in an organized form, for example in rows and columns.
   No of rows called Cardinality and No of columns called Degree of a relation
   Sources: Database, Spread sheets, OLTP systems.

Working with Structured data:

- Storage: Data types – both defined and user defined help with the storage of structured data
- update, delete: Updating, deleting, etc. is easy due to structured form
- Security: can be provided easily in RDBMS.
- Indexing /Searching: Data can be indexed based not only on a text string but other attributes as well. This enables streamlined search
- Scalability (horizontal/vertical): Scalability is not generally an issue with increase in data as resources can be increased easily.
- Transaction Processing (Atomicity, Consistency, Integrity, Durability)

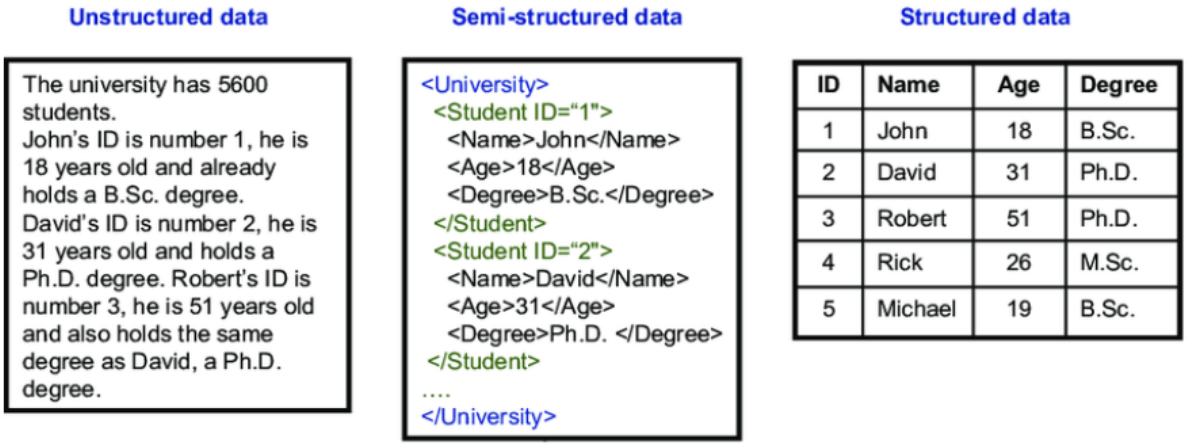

Fig. Sample representation of types of digital data

2. Semi-Structured Data:

This data which doesn't conform to a data model but has some structure. Metadata for this data is available but is not sufficient.
Sources: XML, JSON, E-mail

Characteristics:
- inconsistent structure.
- self describing (label/value pairs)
- schema information is blended with data values
- data objectives may have different attributes not known before

Challenges:
- Storage cost: Storing data with their schemas increases cost
- RDBMS: Semi-structured data cannot be stored in existing RDBMS as data cannot be mapped into tables directly
- Irregular and partial structure: Some data elements may have extra information while others none at all
- Implicit structure: In many cases the structure is implicit.
- Interpreting relationships and correlations is very difficult
- Flat files: Semi-structured is usually stored in flat files which are difficult to index and search
- Heterogeneous sources: Data comes from varied sources which is difficult to tag and search.

3. Unstructured Data:

- This is the data which does not conform to a data model or is not in a form which can be used easily by a computer program.
- About 80–90% data of an organization is in this format.
- **Sources:** memos, chat-rooms, PowerPoint presentations, images, videos, letters, researches, white papers, body of an email, etc.

Characteristics:
- Does not confirm to any data model
- Can't be stored in the form of rows and columns
- Not in any particular format or sequence
- Not easily usable by the program
- Doesn't follow any rule or semantics

Challenges:

- Storage space: Sheer volume of unstructured data and its unprecedented growth makes it difficult to store. Audios, videos, images, etc. acquire huge amount of storage space
- Scalability: Scalability becomes an issue with increase in unstructured data
- Retrieve information: Retrieving and recovering unstructured data are cumbersome
- Security: Ensuring security is difficult due to varied sources of data (e.g. e-mail, web pages)
- Update/delete: Updating, deleting, etc. are not easy due to the unstructured form
- Indexing and Searching: Indexing becomes difficult with increase in data.
- Searching is difficult for non-text data
- Interpretation: Unstructured data is not easily interpreted by conventional search algorithm
- Tags: As the data grows it is not possible to put tags Manually
- Indexing: Designing algorithms to understand the meaning of the document and then tag or index them accordingly is difficult.

Dealing with Unstructured data:

- **Data Mining:** Knowledge discovery in databases, popular Mining algorithms are Association rule mining, Regression Analysis, and Collaborative filtering
- **Natural Language Processing:** It is related to HCI. It is about enabling computers to understand human or natural language input.
- **Text Analytics:** Text mining is the process of gleaning high quality and meaningful information from text. It includes tasks such as text categorization, text clustering, sentiment analysis and concept/entity extraction.
- **Noisy text analytics:** Process of extraction structured or semi-structured from noisy unstructured data such as chats, blogs, wikis, emails, Spelling mistakes, abbreviations, uh, hm, non standard words.
- **Manual Tagging with meta data**: This is about tagging manually with adequate meta data to provide the requisite semantics to understand unstructured data.
- **Parts of Speech Tagging:** POST is the process of reading text and tagging each word in the sentence belonging to particular parts of speech such as noun, verb, objective.
- **Unstructured Information management architecture:** Open source platform from IBM used for real time content analytics.

Q) Define Big Data. What are the characteristics of Big Data?

Big Data is high-volume, velocity, and variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making.

Characteristics(V's):

1. **Volume:** It refers to the amount of the data. The size of the data is being increased from Bits to Yottabytes.

Bits-> Bytes-> KBs-> MBs-> GBs-> TBs-> PBs-> Exabytes-> Zettabytes-> Yottabytes

There are different sources of data like doc, pdf, YouTube, a chat conversation on internet messenger, a customer feedback form on an online retail website, CCTV coverage and weather forecast.

The sources of Big data:

1. Typical internal data sources: data present within an organization's firewall.

Data storage: File systems, SQL (RDBMSs- oracle, MS SQL server, DB2, MySQL, PostgreSQL etc.) NoSQL, (MangoDB, Cassandra etc) and so on.
Archives: Archives of scanned documents, paper archives, customer correspondence records, patient's health records, student's admission records, student's assessment records, and so on.

2. External data sources: data residing outside an organization's Firewall.

Public web: Wikipedia, regulatory, compliance, weather, census etc.,

3. Both (internal + external sources)

Sensor data, machine log data, social media, business apps, media and docs.

2. **Variety:** Variety deals with the wide range of data types and sources of data. Structured, semi-structured and Unstructured.

Structured data: From traditional transaction processing systems and RDBMS, etc.
Semi-structured data: For example Hypertext Markup Language (HTML), eXtensible Markup Language (XML).
Unstructured data: For example unstructured text documents, audios, videos, emails, photos, PDFs , social media, etc.

3. **Velocity:** It refers to the speed of data processing. we have moved from the days of batch processing to Real-time processing.

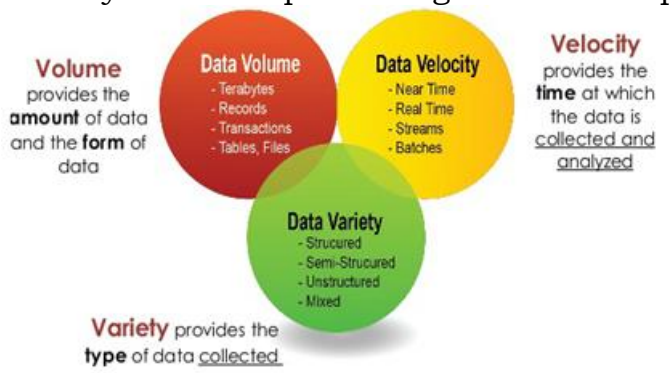

Fig. 3 V's of Big Data

Another V's in Big Data they are

4. **Veracity:** Veracity refers to biases, noise and abnormality in data. The key question is "Is all the data that is being stored, mined and analysed meaningful and pertinent to the problem under consideration".
5. **Value:** This refers to the value that big data can provide, and it relates directly to what organizations can do with that collected data. It is often quantified as the potential social or economic value that the data might create.
6. **Volatility:** It deals with "How long the data is valid? "

7. **Validity:** Validity refers to accuracy & correctness of data. Any data picked up for analysis needs to be accurate.
8. **Variability:** Data flows can be highly inconsistent with periodic peaks.

Q) **How is traditional BI environment different from Big data environment?**

Traditional BI Environment	Big Data Environment
Data is stored in central server	Data is stored in a distributed file system.
Server scales vertically	Distributed file system scales by horizontally.
Analyzes offline or historic data	Analyzes real or streaming data
Supports Strucuted data only	Supports variety of data ie; structured, semi-structured and unstructured data.

Q) Explain evolution of Big Data. What are the challenges of Big Data?

Evolution:

	Data Generation and storage	**Data Utilization**	**Data Driven**
Complex and unstructured			Structured data, Unstructured data, Multimedia data
Complex and Relational		Relational databases : Data intensive applications	
Primitive and structured	Main frames: Basic data storage		
	1970s and before	Relational 1980s and 1990s	2000s and beyond

The challenges with big data:

1. Data today is growing at an exponential rate. Most of the data that we have today has been generated in the last two years. The key question is : will all this data be useful for analysis how will separate knowledge from noise.
2. How to host big data solutions outside the world.
3. The period of retention of big data.
4. Dearth of skilled professionals who possess a high level of proficiency in data science that is vital in implementing Big data solutions.

5. Challenges with respect to capture, curation, storage, search, sharing, transfer, analysis, privacy violations and visualization.
6. Shortage of data visualization experts.
7. Scale : The storage of data is becoming a challenge for everyone.
8. Security: The production of more and more data increases security and privacy concerns.
9. Schema: there is no place for rigid schema, need of dynamic schema.
10. Continuous availability: How to provide 24X7 support
11. Consistency: Should one opt for consistency or eventual consistency.
12. Partition tolerant: how to build partition tolernant systems that can take of both hardware and software failures.
13. Data quality: Inconsistent data, duplicates, logic conflicts, and missing data all result in data quality challenges.

Q) Define Big Data Analytics. What are the various types of analytics?

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.
Few Top Analytics tools are: MS Excel, SAS, IBM SPSS Modeler, R analytics, Statistica, World Programming Systems (WPS), and Weka.
The open source analytics tools are: R analytics and Weka.

Big Data Analytics is:

1. Technology enabled analytics: The analytical tools help to process and analyze big data.
2. About gaining a meaningful, deeper, and richer insights inot business to drive in right direction, understanding the customer's demographics, better leveraging the services of vendors and suppliers etc.
3. About a competitive edge over the competitors by enabling with finding that allow quicker and better decision making.
4. A tight handshake between 3 communities: IT, Business users and Data Scientists.
5. Working with datasets whose volume and variety exceed the current storage and processing capabilities and infrastructure of the enterprise.
6. About moving code to data. This makes perfect sense as the program for distributed processing is tiny compared to the data.

Classification of Analytics: There are basically two schools of thought:

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

First school of thought:

1. Basic analytics: This primarily slicing and slicing 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 process.
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.

Types of Big Data Analytics				
	Descriptive	**Diagnostic**	**Predictive**	**Prescriptive**
Answers the question...	What happened?	Why did it happen?	What will happen next?	What should I do?
Level of advancement	Low	Medium	High	Very high
Incorporates AI and machine learning?	Not usually	Sometimes	Usually	Always
Level of popularity	Used by almost all organizations	Used by many organizations	Used by a smaller but growing group of organizations	Not yet widespread

Second school of thought:

Analytics 1.0	**Analytics 2.0**	**Analytics 3.0**
Era: 1950s to 2009	Era: 2005 to 2012	Era: 2012 to present
Descriptive statistics (report events, occurrences etc of the past.	Descriptive statistics + Predictive statistics (use data from the past to make predictions for the future.	Descriptive statistics + Predictive statistics + prescriptive statistics (use data from the past to make prophecies for the future and at the same time make recommendations to leverage the situations to one's advantage.
Key questions asked: What happened? Why did it happen?	Key questions are: What will happen? Why will it happen?	Key questions are: 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 third party applications.	Big Data	A blend of big data and data from legacy systems, ERP,CRM and third 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 higher pace. This fast flow of big volume data had to be stored and processed rapidly, often on massively 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 being both internally and externally sourced.
Relational databases	Database applications, Hadoopo clusters, SQL to hadoop environments etc..	In ,memory analytics, in database processing, agile analytical methods, Machine learning techniques etc ..

Q) What are the advantages of Big Data Analytics?

- **Business Transformation** In general, executives believe that big data analytics offers tremendous potential to revolution their organizations.
- **Competitive Advantage** According survey 57 percent of enterprises said their use of analytics was helping them achieve competitive advantage, up from 51 percent who said the same thing in 2015.
- **Innovation** Big data analytics can help companies develop products and services that appeal to their customers, as well as helping them identify new opportunities for revenue generation.
- **Lower Costs** In the New Vantage Partners Big Data Executive Survey 2017, 49.2 percent of companies surveyed said that they had successfully decreased expenses as a result of a big data project.
- **Improved Customer Service** Organizations often use big data analytics to examine social media, customer service, sales and marketing data. This can help them better gauge customer sentiment and respond to customers in real time.
- **Increased Security** Another key area for big data analytics is IT security. Security software creates an enormous amount of log data.

Q) List what Big Data Analytics is not?

Big Data Analtics coexist with both RDBMS and Data Warehouse, leveraging the power of each to yield business value.
Big Data Analtics isn't:

- 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
- "One-size fit all" traditionaly RDBMS built on shared disk and memory.

Q) Explain different Big Data Analytics Approaches.

Reactive – Business Intelligence: It is about analysis of the pas or historical data and then displaying the finding of the analysis or reports in the form of enterprise dash boards, alerts, notifications etc.
Reactive – BigData Analytics: Here the analysis is done on huge datasets but the approach is still reactive as it is still base don static data.
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 bigdata as it still used traditional data base management practices.
Proactive – Big Data Analytics: This is sieving through terabytes 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 sovle complex problems using more data.

Q) Explain the following terminology of Big Data

a. In-Memory Analytics
b. In-Database processing
c. Symmetric Mulit-processor system
d. Massively parallel processing
e. Shared nothing architecture
f. CAP Theorem

In-memory Analytics: Data access from non-volatile storage such as hard disk is a slow process. 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.

In-Database Processing: In-Database processing is also called In-database analytics. It works by fusing data warehouses with analytical systems. Typically the data from various enterprise OLTP systems after cleaning up through the process of ETL is stored in the Enterprise Dataware house or data marts. The huge data sets are then exported to analytical programs for complex and extensive computations.

Symmetric Multi-Processor System:
In this there is 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 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.

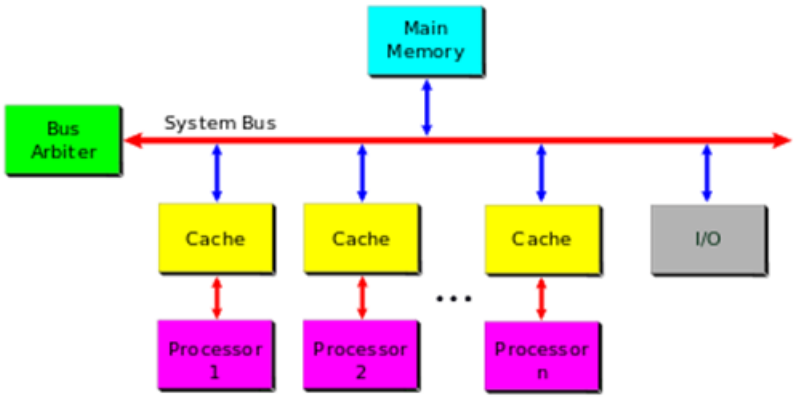

Fig. Symmetric Multiprocessor System(SMP)

Massively Parallel Processing:
Massively parallel Processing (MPP) refers to the coordinated processing of programs by a number of processors working parallel. The processors each have their own OS and dedicated memory. They work on different parts of the same program. The MPP processors communicate using some sort of messaging interface.
MPP is different from symmetric multiprocessing in that SMP works with processors sharing the same OS and same memory. SMP also referred as tightly coupled Multiprocessing.

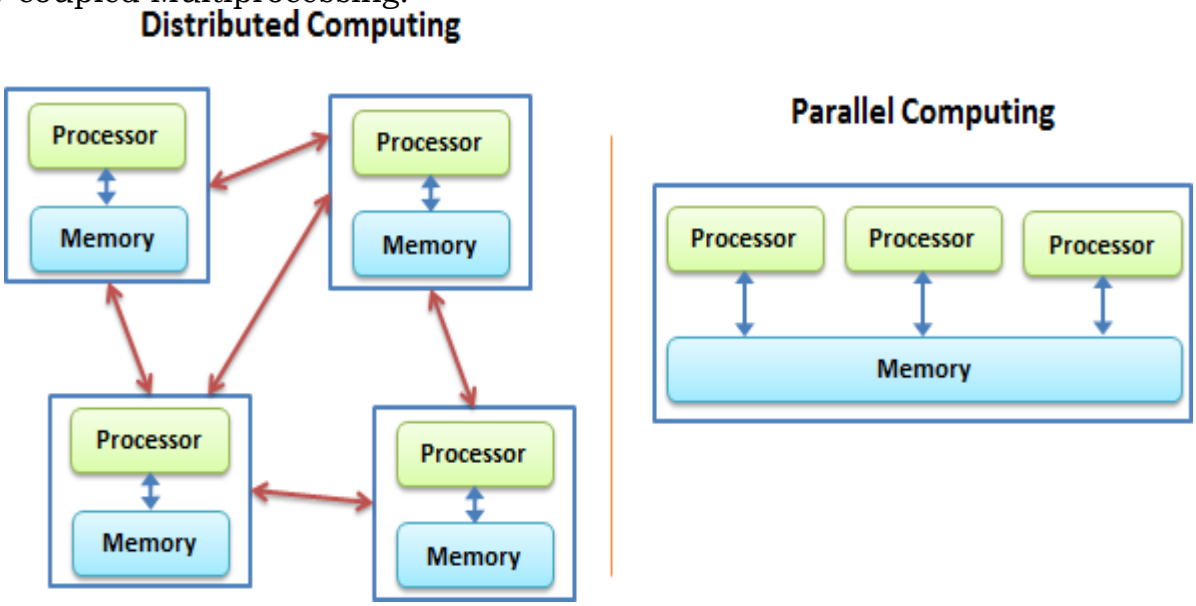

Fig. Distributed Computing an Parallel Computing Environments

Shared nothing Architecture: The three most common types of architecture for multiprocessor systems:

1. Shared memory
2. Shared disk
3. Shared nothing.

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.

Advantages of shared nothing architecture:
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. Scalability: Assume that the disk is a shared resource it implies that the controller and the disk band-width 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 the scalability.

CAP Theorem: The CAP theorem is also called the Brewer's theorem. It states that in a distributed computing environment, it is impossible to provide the following guarantees. At best you can have two of the following three and one must be sacrificed.

1. Consistency
2. Availability
3. Partition tolerance

1. Consistency implies that every read fetches the last write. Consistency means that all nodes see the same data at the same time. If there are multiple replicas and there is an update being processed, all users see the update go live at the same time even if they are reading from different replicas.
2. Availability implies that reads and writes always succeed. Availability is a guarantee that every request receives a response about whether it was successful or failed.
3. Partition tolerance implies that the system will continue to function when network partition occurs. It means that the system continues to operate despite arbitrary message loss or failure of part of the system.

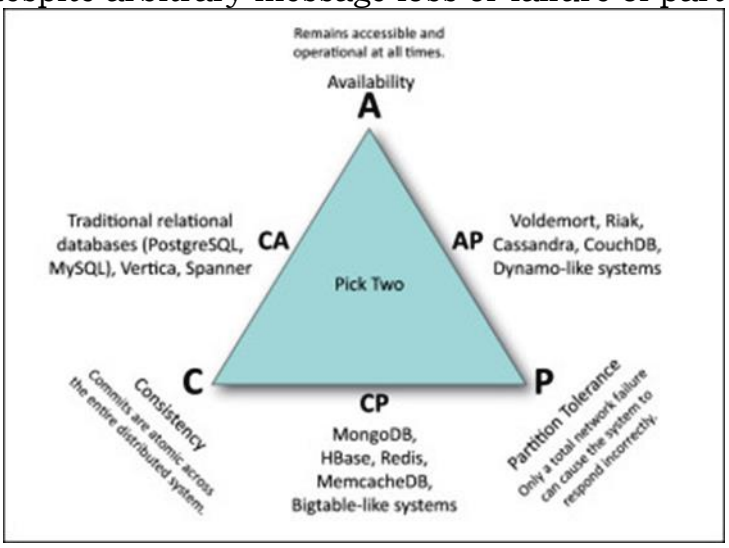

Fig. Databases and CAP

Q) What is BASE?

Basically Available, Soft State, Eventual Consistency (BASE) is a data system design philosophy that in distributed environment, it gives importance to availability over consistency of operations.
BASE may be explained in contrast to another design philosophy - Atomicity, Consistency, Isolation, and Durability (ACID). The ACID model promotes consistency over availability, whereas BASE promotes availability over consistency.

Q) Give Real-time applications of Big Data Analytics.

1. **Banking and Securities Industry:**

- This industry also heavily relies on Big Data for risk analytics, including; anti-money laundering, demand enterprise risk management, "Know Your Customer," and fraud mitigation.
- The Securities Exchange Commission (SEC) is using Big Data to monitor financial market activity. They are currently using network analytics and natural language processors to catch illegal trading activity in the financial markets.

2. **Communications, Media and Entertainment Industry:**

Organizations in this industry simultaneously analyze customer data along with behavioral data to create detailed customer profiles that can be used to:

- Create content for different target audiences
- Recommend content on demand
- Measure content performance

Eg.

- **Spotify,** an on-demand music service, uses Hadoop Big Data analytics, to collect data from its millions of users worldwide and then uses the analyzed data to give informed music recommendations to individual users.
- **Amazon Prime**, which is driven to provide a great customer experience by offering video, music, and Kindle books in a one-stop-shop, also heavily utilizes Big Data.

3. **Healthcare Sector:**

- Some hospitals, like Beth Israel, are using data collected from a cell phone app, from millions of patients, to allow doctors to use evidence-based medicine as opposed to administering several medical/lab tests to all patients who go to the hospital.
- Free public health data and Google Maps have been used by the University of Florida to create visual data that allows for faster identification and efficient analysis of healthcare information, used in tracking the spread of chronic disease.

4. **Education:**

- The University of Tasmania, An Australian university with students has deployed a Learning and Management System that tracks, among other things, when a student logs onto the system, how much time is spent on different pages in the system, as well as the overall progress of a student over time.
- On a governmental level, the Office of Educational Technology in the U. S. Department of Education is using Big Data to develop analytics to help correct course students who are going astray while using online Big Data certification courses. Click patterns are also being used to detect boredom.

5. **Government:**

- In public services, Big Data has an extensive range of applications, including energy exploration, financial market analysis, fraud detection, health-related research, and environmental protection.

- The Food and Drug Administration (FDA) is using Big Data to detect and study patterns of food-related illnesses and diseases.

6. Insurance Industry:

Big data has been used in the industry to provide customer insights for transparent and simpler products, by analyzing and predicting customer behavior through data derived from social media, GPS-enabled devices, and CCTV footage. The Big Data also allows for better customer retention from insurance companies.

7. Transportation Industry:

Some applications of Big Data by governments, private organizations, and individuals include:

- Governments use of Big Data: traffic control, route planning, intelligent transport systems, congestion management (by predicting traffic conditions)
- Private-sector use of Big Data in transport: revenue management, technological enhancements, logistics and for competitive advantage (by consolidating shipments and optimizing freight movement)

8. Energy and Utility Industry:

Smart meter readers allow data to be collected almost every 15 minutes as opposed to once a day with the old meter readers. This granular data is being used to analyze the consumption of utilities better, which allows for improved customer feedback and better control of utilities use.

Q) What is NoSQL? What is the need of NoSQL? Explain different types of NoSQL databases.

NoSQL Stands for **N**ot **O**nly **SQL.** These are non-relational, open source, distributed databases.

Features of NoSQL:

1. NoSQL databases are non-relational: They do not adhere to relational data model. In fact either key-value pairs or document oriented or column oriented or graph based databases.
2. Distributed: The data is distributed across several nodes in a cluster constituted of low commodity hardware.
3. No Support for ACID properties: They do not offer support for ACID properties of transactions. On the contrary, they adherence to CAP theorem.
4. No fixed table schema: NoSQL databases are becoming increasing popular owing to their support for flexibility to the schema. They do not mandate for the data to strict adhere to any schema structure at the time of storage.

Need of NoSQL:

1. It has scale out architecture instead of the monolithic architecture of relational databases.
2. It can house large volumes of structured, semi-structured and unstructured data.
3. Dynamic Schema: It allows insertion of data without a predefined schema.

4. Auto Sharding: It automatically spread data across an arbitrary numer of servers or nodes in a cluster.
5. Replication: It offers good support for replication which in turn guarantees high availability, fault tolerance and disaster recovery.

Types of NoSQL databases: They broadly divided into Key-Value or big hash table and Schemal-less.

1. **Key-Value:** It maintains a big hash table of keys and values.
   Key are unique.
   It is fast, scalable and fault tolerance.
   It can't model more complex data structure such as objects
   Eg. Dynamo, Redis, Riak etc.
   Sample Key-Value pair database:

Key	Value
Fname	Praneeth
Lname	Ch

2. **Document:** It maintains data in collections constituted of documents.
   Eg. MongoDB, Apace CouchDB, Couchbase, MarkLogic etc.
   Sample Document in Document DB:

```
{
        "Book Name": "Big Data and Analytics",
        "Publisher": "Wiley India",
        "Year": "2015"
}
```

3. **Column:** Each storage block has data from only one column. It only fetch column families of those columns that are required by a query (all columns in a column family are stored together on the disk, so multiple rows can be retrieved in one read operation à data locality
   Eg. Cassandra, HBase etc.

   Sample column database:
   UserProfile = {
   Cassandra = { emailAddress:"casandra@apache.org" , age:"20"}
   TerryCho = { emailAddress:"terry.cho@apache.org" , gender:"male"}
   Cath = { emailAddress:"cath@apache.org" ,
   age:"20",gender:"female",address:"Seoul"}
   }

4. **Graph:** They are also called Network database. A graph stores data in nodes.
   Data model:
   - (Property Graph) nodes and edges
     - Nodes may have properties (including ID)
     - Edges may have labels or roles
   - Key-value pairs on both

   Eg. Neo4j, HyperGraphDB, InfiniteGraph etc.
   Sample Graph database:

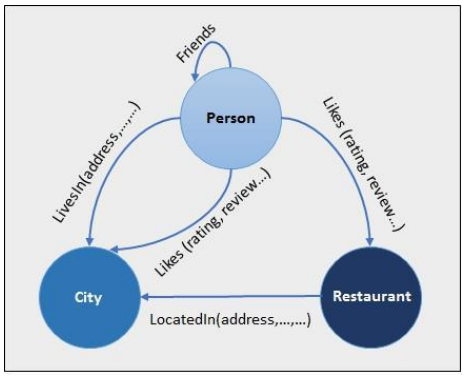

Fig. Sample Graph Database

Q) What are the advantages and disadvantage of NoSQL?

Advantages:

- Big Data Capability
- No Single Point of Failure
- Easy Replication
- It provides fast performance and horizontal scalability.
- Can handle structured, semi-structured, and unstructured data with equal effect
- NoSQL databases don't need a dedicated high-performance server
- It can serve as the primary data source for online applications.
- Excels at distributed database and multi-data centre operations
- Eliminates the need for a specific caching layer to store data
- Offers a flexible schema design which can easily be altered without downtime or service disruption

Disadvantages:

- Limited query capabilities
- RDBMS databases and tools are comparatively mature
- It does not offer any traditional database capabilities, like consistency when multiple transactions are performed simultaneously.
- When the volume of data increases it is difficult to maintain unique values as keys become difficult
- Doesn't work as well with relational data
- Open source options so not so popular for enterprises.
- No support for join and group-by operations.

Q) Differentiate SQL and NoSQL.

SQL	NoSQL
Relational database	Non-relational, distributed database
Relational model	Model-less approach
Pre-defined schema	Dynamic schema for unstructured data
Table based databases	Document-based or graph-based or wide column store or key-value pairs databases

Vertically scalable (by increasing system resources)	Horizontally scalable (by creating commodity machines)
Uses SQL	Uses UnQL (Unstructured Query Language)
Not preferred for large datasets	Largely preferred for large datasets
Not a best fit for hierarchical data	Best fit for hierarchical storage as it follows the key-value pair of storing data similar to JSON (Java Script Object Notation)
Emphasis on ACID properties	Follows Brewer's CAP theorem
Excellent support from vendors	Relies heavily on community support
Supports complex querying and keeping needs	Does not have good support for complex querying
Can be configured for strong consistency	Few support strong consistency (e.g., MongoDB), few others can be configured for eventual consistency (e.g., Cassandra)
Examples: Oracle, DB2, MySQL, MS SQL, PostgreSQL, etc.	MongoDB, HBase, Cassandra, Redis, Neo4j, CouchDB, Couchbase, Riak, etc.

Q) Explain where to use NoSQL? Explain some real time applications of NoSQL.

Key-Value
- Shopping carts
- Web user data analysis
- Amazon, Linkedin

Document based
- Real-time Analysis
- Logging
- Document archive management

Column-oriented
- Analyze huge web user actions
- Sensor feeds
- Facebook, Twitter, eBay, Netfix

Graph-based
- Network modeling
- Recommendation
- Walmart-upsell, cross-sell

Real time applications of NoSQL in BigData Analytics:

- HBase for Hadoop, a popular NoSQL database is used extensively by Facebook for its messaging infrastructure.
- HBase is used by Twitter for generating data, storing, logging, and monitoring data around people search.

- HBase is used by the discovery engine Stumble upon for data analytics and storage.
- MongoDB is another NoSQL Database used by CERN, a European Nuclear Research Organization for collecting data from the huge particle collider "Hadron Collider".
- LinkedIn, Orbitz, and Concur use the Couchbase NoSQL Database for various data processing and monitoring tasks.

Q) What is NewSQL? Differentiate SQL, NoSQL and NewSQL

NewSQL supports relational data model and uses SQL as their primary interface.
NewSQL Characterisitcs:

- SQL interface for application interaction
- ACID support for transactions
- An architecture that provides higher per node performance vis-a-vs traditional RDBMS solution
- Scale out, shared nothing architecture
- Non-locking concurrency control mechanism so that real time reads will not conflict with writes.

	SQL	NoSQL	NewSQL
Adherence to properties	Yes	No	Yes
OLTP/OLAP	Yes	No	Yes
Schema rigidity Adherence to data model	Yes Adherence	No	Maybe
Data Format Flexibility	No	Yes	Maybe
Scalability	Scale up Vertical Scaling	Scale out Horizontal Scaling	Scale out
Distributed Computing	Yes	Yes	Yes
Community Support	Huge	Growing	Slowly growing