UNIT-1

- **Big Data** → 3V → high volume, high velocity, high variety → demands → cost effective, innovative forms of information processing for better decision.

→ Black Box – Planes / Helicopters
→ Social Media Data → Search Engine Data
→ Stock Exchange Data

- **Types of Big Data**

1. **Structured Data** — organized data, predefined format, defined repeating pattern. Data stored in fixed field in a record / file. Entities & their attributes are mapped.
   - Sources: Comp / Machine Generated → sensor data – GPS / RFID; weblog / financial data; stock market; Point of sale data → barcode, cardswipe. Human Generated → gaming related, in interaction with comp, I/P data, e.g. click-stream – coherence at click a link.
2. **Unstructured Data** — No logical, repeating pattern. Different formats – email / text / audio / video. Sources – Mobile / Social Media.
   - Machine Generated: satellite image, scientific data, Photograph & video, Radar / Sonar, CCTV.
   - Human Generated – Social Media / Mobile Data / Texts / website content
3. **Semi Structured Data** — Schema less / self describing structure. Data stored inconsistently in rows and columns.

- **Types of sources**

1. **Internal source** — structured / organized data; originates from within enterprise.
   CRM / ERP
2. **External Source** — Unstructured / Unorganized data; originates from external environment of organization.
   Ex: Buisness Partner, Internet / Govt

- **Information System**
  - system used for organizing & processing info. elements of
  - information technology + people's activities using technology to support operations / management

(combination of)

→ Interaction b/w people, algorithmic processes, data and technology.

→ Mediator b/w actions and technology.

- processes
- transmits
- storing & retrieving
- manipulating
- displaying

] Information.

- **Reasons to Use Information System**

1. Operational Excellence
   - ~~improves efficiency of operations~~
   - achieve higher profitability.
   - high level of efficiency & productivity

] can predict customer's choice

2. Improved Decision Making
3. Customer – Supplier Good Relations
3. New Product, Services

Elements of Big Data

1. **Volume** – amount of data. Nowadays → PetaBytes / Exa-byte. By 2020 25 zettabytes.

2. **Velocity** – speed of data processing.
   - velocity at which it is created & integrated.
   - rate at which data is generated & shared

3. **Variety**
   - No. of types of data.
   - data is generated from different types of sources [external / internal]

4. **Veracity** → uncertainity of data
   - whether data obtained in correct or consistent.

• Uses of Big Data

1. **Understanding & Targeting customers.**
   - understand customers & their behaviors.
   - social media data / browse logs / text analytics.

Understanding & Optimizing Buisness Processes.
- optimize their stocks based on predictions.
- predictions → social media data / web search trends.

Personal Qualification & Optimization.
- Benefit from data generated from wearing devices.
  Smart watches.

4. Improving Healthcare & Public Health
   - computing power of big data - decodes entire DNA strand within a minute
   - can find new cures & predict disease patterns

5. Sports Performance Improvement
   - video analytics - tracks performance of every player
   - sensor technology - gives feedback

6. Science & Research Improvement

7. Improving Security & Law Enforcement
   - detect & prevent cyber attacks
   - Big data tools to catch criminals
   - predict criminal activity
   - credit card companies - use big data → detect fraud

Big Data Management - cycle

Capture → Organize → Integrate → Analyze → Act

Big Data Analytics

- process of examining data
  - variety of sources / types
  - volumes / complexities.
  - ↓ to uncover hidden patterns / unknown correlations and other useful information

1. Traditional Big Data Analytics

- enterprise will have computer to store & process data.
- data stored in → RDBMS → oracle DB / MS SQL server ##
- sophisticated software written to interact with DB process data & present to users for analysis purpose.

Limitation
- cannot deal with huge amount of data
- can work well with only less volume of data.

Advanced Big Data Analytics.

- Map Reduce Algo - solves the problem of tradional analyt
  - ↓
  - → divides task into smaller parts.
  - → assign those parts to many computers connected over networ
  - → collects results to form the final result dataset.

- **Distributing Computing**
  - multiple computing resources are connected in a network
  - computing tasks are distributed across this network

Sharing of tasks - increases speed & efficiency of system.

Latency Problem -

Latency
- delay with a system based on delays in execution of a task.
- issue in every aspect of computing
  - communication
  - data manage
  - system
  - Performan

Distributed computing can solve this problem.

Big Data companies - need low latency because of BD requires high speed & volume.

Sources of Big Structured Data

- Computer / Machine Generated
  - Sensor data
  - Web log data
  - Point of Sale
  - Financial data
- Human Generated
  - Input data
  - Click-stream data
  - Game related data

Sources of Big Unstructured Data

- Machine Generated
  - Satellite Images
  - Scientific Data
    - seismic
    - atmospheric.
  - Radar / Sonar
  - Photographic & Video
- Human Generated.
  - Text internal to company - logs / surveys / results / emails.
  - Social media data
  - Mobile Data
  - Website content.

Role of Relational Database in Big Data

Data presistence – how db retains itself when modified.

RDBMS – primitive technique for data presistence.

→ data is stored in table.

→ have a schema – structural representation of what is in the db
defines table, fields in table & relationship b/w two.

→ Data Stored in rows & columns.

→ SQL can be used to retrive data.

- Role of CMS IN Big Data Management ;

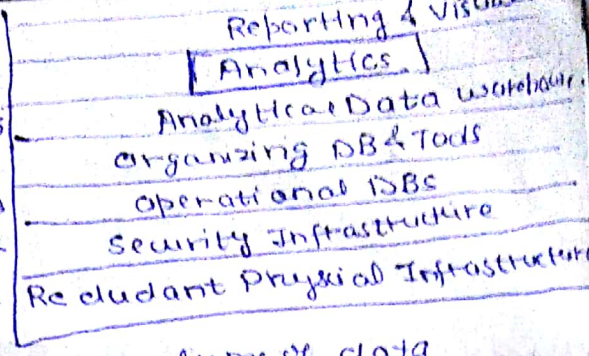
CMS – content Management System

- can manage complete lifecycle of content
  web content / document content & other forms media
- comprises – strategies / methods / tools.
  ↓
  capture data
  manage & store, preserve.
  & deliver content & documents
- Include technologies
  - document management
  - web content management
  - collaboration.

document content — collaboration — web content

Big Data Stack

UNIT-2



1. Redundant Physical Infra
   - fundamental to scalibility & operation of BD architecture.
   - supports unanticipated / unpredictable volume of data
2. Security Infrastructure
   - more important is BD analysis, more important is to secure data
3. Operational Data sources (Structured / Unstructured / semistructured)
   - incoperate all the data sources to give complete picture of business
   - highly structured data managed in relational database.

---

4. REDUNDANT PHYSICAL INFRASTRUCTURE
- Network Design
- Hardware → Storage & Servers
- well managed environment

1. Physical Redundant Big Data Networks
- Network should be redundant & have enough capacity → accomdate anticipate volume & velocity of inbound & outbound dat
- Designers plan
  - expected increase.
  - physical implementation — elastic.

2. Manage Big Data Hardware : Storage & Servers.
- Hardware — sufficient speed & capacity
  - to handle all expected Big data capabili
- If slow servers → bottleneck in network.
- fast servers — overcome variable N/w performance.

3) Big Data Infrastructure Operations Managment

- better performance and flexibility is only possible in a well managed environment.
- Data center managers → able to anticipate & prevent failures to the integrity of data

→ Identify the type of Data needed for Big Data
→ Exploratory stage for big data — identifying patterns in the data.
→ codifying stage for big data — pattern identification to implementation
→ data integration & incorporation stage — after analysis
  - integrate results to real-time business actions

identify the type — explore the data → codifying → data integration
~~(identify pattern)~~ ~~(pattern identification)~~

2) SECURITY INFRASTRUCTURE

1. Data Access — Only authorized used who have legitimate business needs should examine or interact with data.
   - core data storage platforms — rigirous security schemes.
2. Application Access
   - APIs — offers protection from unauthorized usage or acess
3. Data Encryption
   - most challenging aspect of security in BD
   - Identify data elements requiring more security
     ↓
     encrypt only necessary items
   - to provide more & fast computational capability.
4. Threat Detection
   - mobile devices / social networks exponentially increases both amount of data and security threats
   - Multiperimeter approach to security.

~~(2) Opera~~

(3) Operational Databases

core of big data environments → database engines having a collection of data
→ Needs to be fast / scalable / Rock Solid

A - Atomicity → all or nothing - If any part of transcation fails - entire transcation fails.
C - consistency → Transcations only with valid data will be performed.
I - Isolation → Multiple, simultaneous transcations will not interfere.
D - Durability → After the data from transcation is written to db
↳ it stays forever.

- Keeps track of critical data required for real time, day to day operation.
- continuously updated based on transcations happening.
- Able to scale up - to support thousand of users.
- accurate representation of buisness - must blend structured & unstructured data.

~~(4) Organizing Database Services Tools.~~

(4) Organizing Data Services and Tools.

↳ capture, validate and assemble various big data elements into relevant collections.

Organizing Data Services → Map Reduce Engines.
↓
Designed to optimize the organization of big data.

- ecosystem of tools & technologies
  - ↳ used to gather & assemble data in preparation for further processing.
- Tools provide
  - integration.
  - translation.
  - normalization.

⑤ Analytical Data warehouses

Data warehouses – contain normalized data gathered from variety of sources & assembled to optimize decision making

- simplify creation of reports & visualization of disparate data items.

⑥ Big Data Analytics

Analytics Tools & techniques – helpful in making sense of big data.
- Able to work with large amount of potentially real time data.

Reporting & Dashboards – user friendly representation of info from various sources

Visualization – Next step in evolution of reporting
- O/P – tends to be highly interactive & dynamic in nature.
- Difference b/w reports & visualized O/P → animation.

Analytics & Advanced Analytics
- Reaching to data warehouse & process the data for human consumption.

VIRTUALIZATION

- Process of using comp resources to imitate other resources.
- Increases
  - IT resource utilization
  - efficiency
  - scalability
  - Improves Latency.

Rather than assigning a set of dedicated physical resources to each task, a pool of virtual resources can be allocated as per need.

Application can use a resource without any concern that

- where it resides
- how it is implemented
- which platform it uses.
- How much of it is available.

- Separates resources & services from underlying physical delivery environment.
- Enables to create many virtual machines within a single physical machine.

Advantages

1. Virtualization of physical resources (server/storage / N/w).
   - improves the resource-utilization.
   - control over usage & performance of IT resources.
2. Provides a level of automization & standardization.
3. Provides foundation for cloud computing.

Characteristics - to support scalability & operation efficiency in big data environment

1. Partitioning - many V.Ms can be supported in single physical system by partitioning available resources.
2. Isolation - Each V.M is isolated from its host & other VMs.
   If one VM crashes - others are not affected.
3. Encapsulation - A V.M can be represented & stored as single file.
   Easy identification - based on services provided.

Network Virtualization – software defined networking

- effient way - to use networking as a pool of connection resources.
- Instead of relying on physical N/w for managing traffic b/w connections.
  ↓
  can create multiple virtual networks.
  all utilizing same physical implementation

→ creates a logical software-based view of the h/w and s/w networking devices (switches, routers etc.)

→ Physical N/w devices – simply forward the packet
Virtual N/w – provides an intelligent abstraction that makes it easy to deploy & manage N/w services.

Cost saving / Efficiency / Security / Flexibility

- Processor & Memory Virtualization

Processor Virtualization [helps to optimize processor
maximize performance

Memory Virtualization – decouples memory from servers.

⇒ In Big data analytics – repeated queries of large dataset & creation of advanced analytic algorithms are there, to discover hidden patterns.

Advanced analytics – require lots of processing power (CP and memory (RAM).. without sufficient CPU & memory these computation can take long time.

processor & memory virtualization can help speed the processing & get analysis results soonel

Server Virtualization → increases efficiency / comp...

- one physical server is partitioned into many virtual servers
- H/w & resources → RAM / CPU / Hard drive can be virtualized
- Each virtual machine → has its own O.S & run its applications
  - software representation of physical machine

Hypervisor → Technology that manages traffic b/w VM & physical machi...

Server Virtualization uses Hypervisor → provide efficiency in the use of physical resources
- ensures that platform can scale up as needed to handle large volume of data as we don't know the amount of data before beginning of analysis.

FOUNDATION THAT ENABLES MANY CLOUD SERVICES USES AS.
DATA SOURCE IN BIG DATA ANAYSIS.

Application Virtualization

- efficient way to manage applications in context with customer demand
- Application is encapsulated
  - in a way that removes dependencies from physical comp. system
- Improves overall manageability & portability of application

Data & Storage Virtualization

Data Virtualization – can create create a platform for dynamic linked data services
- allows data to be easily searched & linked through a unified reference source.

→ It provides an abstract service that delivers data in a consistent form regardless of underlying physical database
→ Exposes ~~data~~ cached data to an application to improve performance

Storage Virtualization – combines physical storage resources so that are more effectively shared.
- reduces cost of storage
- makes it easier to manage data for analysis.

Some data may be unstructured — not easily stored using traditional methods.
↓
SV makes it easier to store large & unstructured data.

DV + SV → makes it easier, less costly
↓
to retrieve & analyze large volume of data

Managing Virtualization with Hypervisor

Hypervisor
- technology responsible for ensuring that resource sharing takes place in an orderly & repeatable way.
- allows multiple O.S to share a single host.
- Creates and run virtual machines. — practical way of getting things virtualized quickly efficiently.

→ It seems like every O.S has physical resources all to itself.
→ In big data, one has to support multiple O.S - hypervisor. becomes ideal delivery mechanism for it.
→ It lets you share same application on multiple systems without having to physically copy that application on each system

Each virtual machine running on a physical machine → Guest Machine

Hypervisor
- schedules the access to guest O.S to have everything. including CPU, memory, disk I/O etc.
- It does all the heavy lifting
  - ↓
  - Guest O.S does not have any idea that it's running in virtual ~~partition~~ partition.

Type 1 Hypervisor
- runs directly on hardware platform.
- achieve higher efficiency - running directly on platform

Type 2 Hypervisor
- Runs on host operating System
- Used - when a need exists to support broad range of I/O devices.