Fundamentals of Data Engineering

Trainer: Pradnyaa S. Dindorkar

BigData => section B => 5 questions

1day ->

2day

3day -> Q1   10
4day -> Q2   10
5 day -> End module 20
 20

per - requistes

1: appln dev / programming
2: Database
3: networking
4: OS - Linux

PG-DBDA

Introduction -syllabus

- **Big Data Fundamentals**
  - Evolution of Data Enggineering | V's: Volume, Velocity, Variety, Veracity, Value
- **Databases**
  - RDBMS - ACID, SQL (basic concept only) | NoSQL - BASE, CAP theorem
- **Data warehouse - OLAP vs OLTP**
  - Data cleansing, Data transformations and Data modelling | Data warehouse vs Data mart
- **Data Engineering Life Cycle**
  - Source → Ingestion → Storage → Transformation → Serving
  - Ingestion: ETL vs ELT
  - Storage: Distributed storage, Storage services | Processing: Batch vs Stream
- **Cloud computing fundamentals**
  - Virtualization, Scaling, Elasticity, Cloud service models, Vendors
- **Big Data Technologies**
  - Frameworks: Hadoop, Hive, Spark, Kafka
  - Applications and Job profiles.

Data Engineering at a Glance

@ 1980

1 - CPU

before 1970 file i/o

Database & Warehouse

@1970
RDBMS   KB

Relational DBMS
C R U D

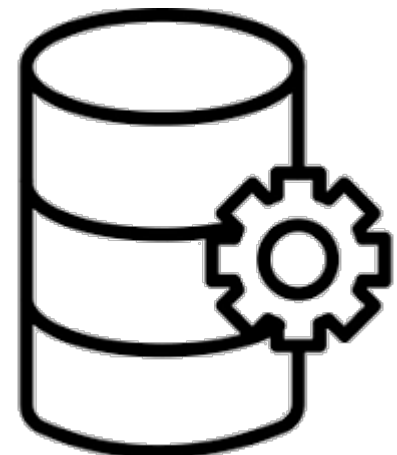@1990 = DWH

Internet & DotCom

1991-1995

Mb+ GB

@2000
data burst

@1998
NoSQL Database

C R U D

4+2

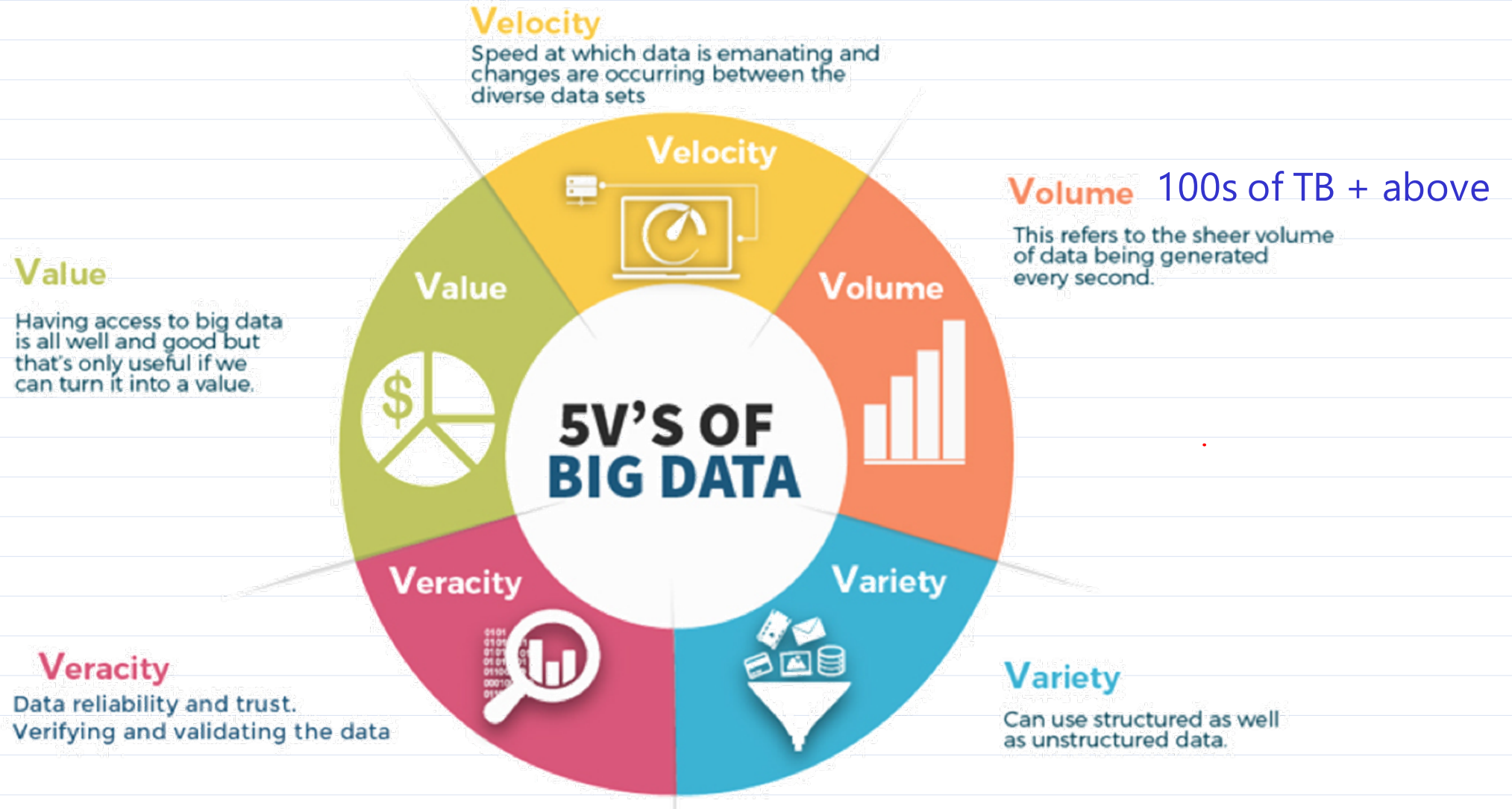massive parallel processing
MPP & Big Data Tech

@2003
TB+PB

Cloud Computing

rent

@2010

Big Data characteristics

100s of TB + above

Types of Data

2008=FB

fixed schema

{
post:----
likes: ----
image:---
png/jpg
gif
}

video:---------
tag:--------
loc:------
feeling:----

STRUCTURED DATA
Uses pre-defined data models filled with labels, numbers and values.

Excel spreadsheets, electronic forms, data tables

SEMI-STRUCTURED DATA
Mainly unstructured but uses internal tags and markings to help classify.

Email stores, JSON, NoSql, XML

UNSTRUCTURED DATA
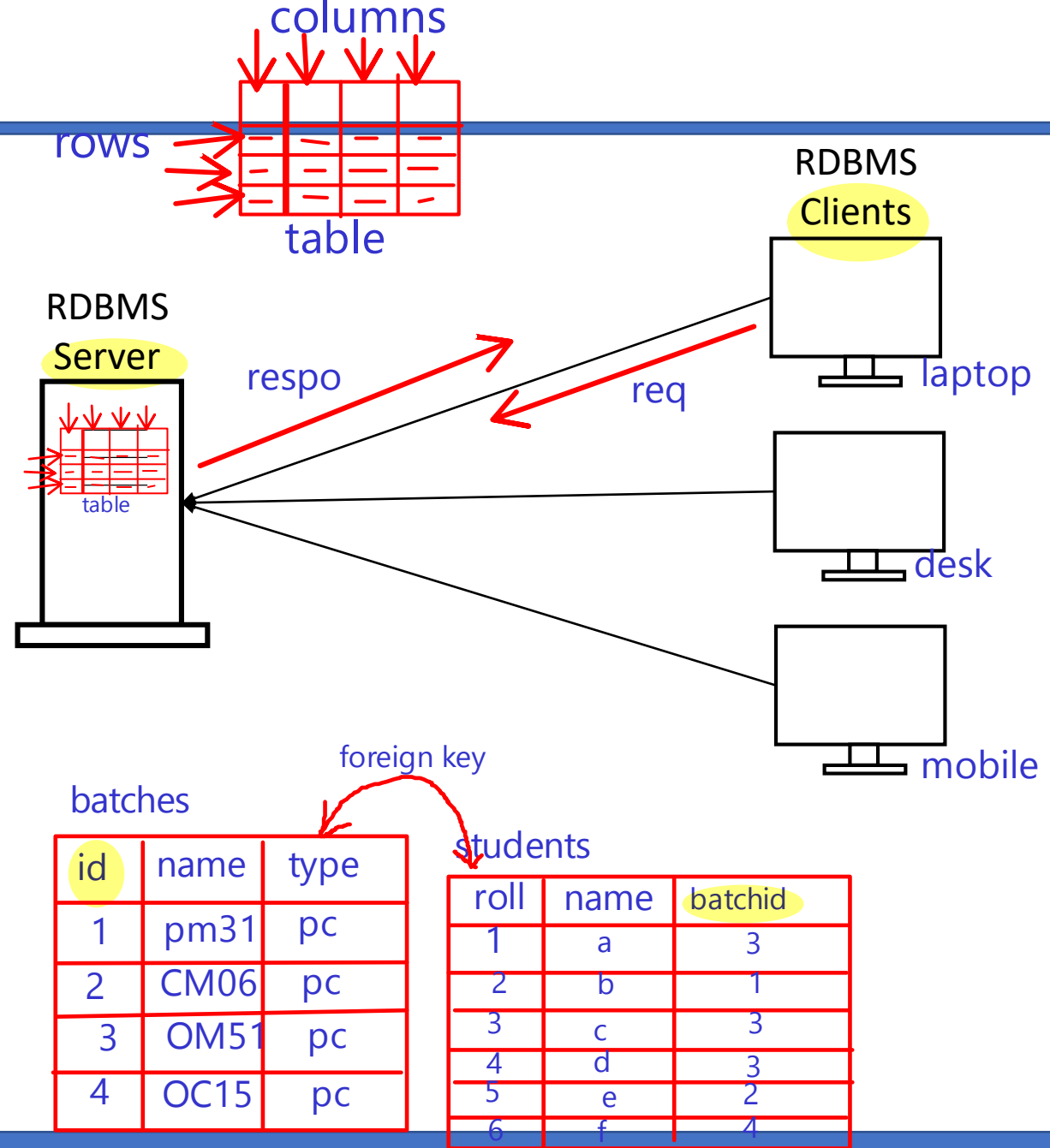No pre-defined data model; packed with text and information.

Scanned PDFs, text documents, audio or video files

images

fixed format schema

flexible schema

ML

RDBMS

- **Every enterprise application need to manage data.**

- **RDBMS is relational DBMS than manages structured data.**

- **Data is organized into tables, rows and columns. Tables are related to each other.**

- **All enterprise RDBMS follow server-client architecture, have built-in relational capabilities, fully ACID transactions, based on Codd's rules.**

- **DB2, Oracle, MS-SQL, MySQL, Postgre-SQL, MS-Access, SQLite, etc.**

columns

rows

table

RDBMS

Server

respo

req

RDBMS

Clients

laptop

desk

mobile

join

foreign key

batches

id	name	type
1	pm31	pc
2	CM06	pc
3	OM51	pc
4	OC15	pc

students

roll	name	batchid
1	a	3
2	b	1
3	c	3
4	d	3
5	e	2
6	f	4

SQL – Structured Query language

- **RDBMS data is processed with SQL queries.**

- **ANSI standardised in 1986 and ISO Standardization in 1987.**

- **Five major categories:**
  - **DDL: Data Definition Language e.g. CREATE, ALTER, DROP, RENAME.**
    - CREATE TABLE student(roll INT, name CHAR(40), batchid INT);
  - **DML: Data Manipulation Language e.g. INSERT, UPDATE, DELETE.**
    - INSERT INTO student VALUES(1, 'Ravi', 3);
    - UPDATE student SET name='Ravee' WHERE roll=1;
    - DELETE FROM student WHERE roll=1;
  - **DQL: Data Query Language e.g. SELECT.**
    - SELECT * FROM student;
  - **DCL: Data Control Language e.g. CREATE USER, GRANT, REVOKE.**
  - **TCL: Transaction Control Language e.g. SAVEPOINT, COMMIT, ROLLBACK.**

Transaction characteristics- ACID

Accounts

id	type	balance
1	save	30000
2	save	5000
3	curr	80000
4	save	45000

4000

Accounts

id	type	balance
1	save	~~30000~~ 26000
2	save	~~5000~~ 9000
3	curr	80000
4	save	45000

30000-4000=26000   acc1

5000+4000=9000   acc2

update 1
update 2

Transaction = > Set of DML queries executed as a single Unit

i:e either all queries in Transaction are successful or
 all queries in Transaction are discarded

Atomic

Consistent =>same result shown to all the clients

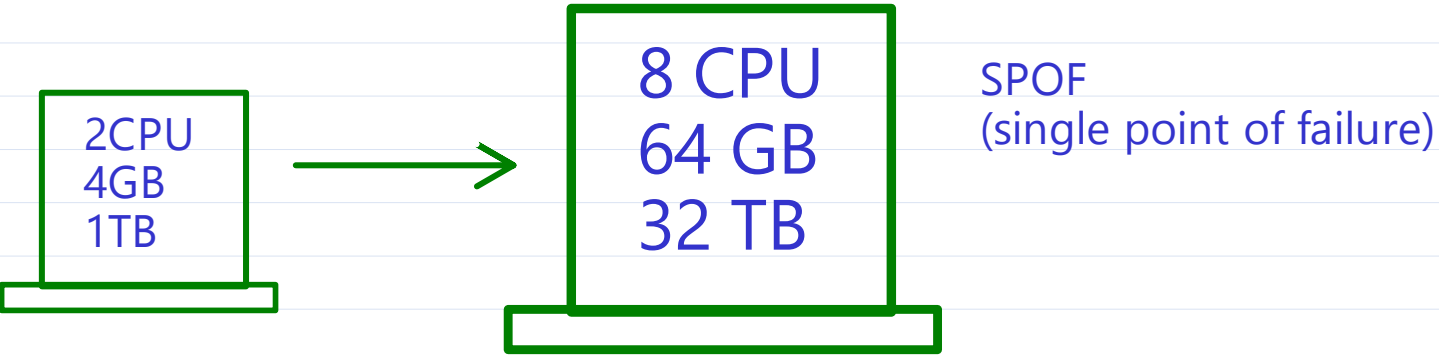Isolated => sevaral Transaction executed simultaneously without affecting each other

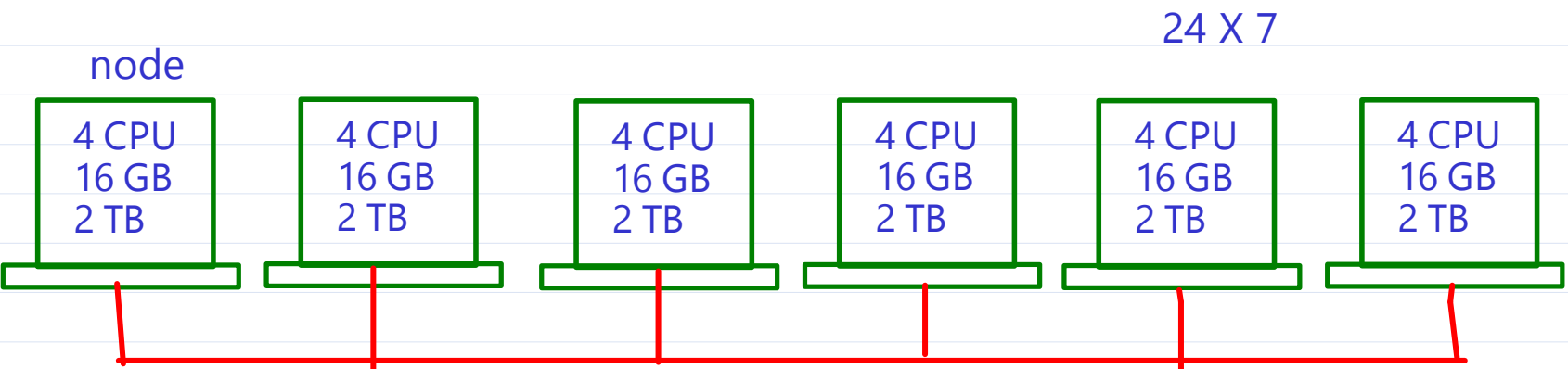Durable=> all changes are saves permanetly

Scalibility

work
clients

data   load

vertical scaling
[up-scaling]

2CPU
4GB
1TB

→

8 CPU
64 GB
32 TB

SPOF
(single point of failure)

24 X 7

node

| 4 CPU
16 GB
2 TB | 4 CPU
16 GB
2 TB | 4 CPU
16 GB
2 TB | 4 CPU
16 GB
2 TB | 4 CPU
16 GB
2 TB | 4 CPU
16 GB
2 TB |

Horizantal scaling
[out-scaling]

Distributed compution

cluster => set of computors connected in a network for dedicated task / work

- Scalability is the ability of a system to expand to meet your business needs.

- Scalability describes a elasticity of the system, ability to adapt to change and demand.

- Good scalability ensures the quality of your service.
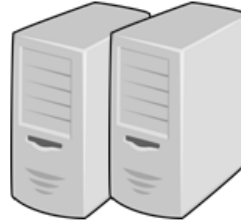
CPU: 3 , RAM: 6G

CPU: 2 , RAM:4G

CPU: 1 , RAM:2G

Vertical scaling describes adding more resources to your current machines.

✓ increase memory in the system

✓ expanding storage by adding hard drives

✓ upgrading the CPUs.

✓ upgrading network speed.

Horizontal Scaling

1 PC (CPU: 1 , RAM:2G)   2 PC (CPU: 1 , RAM:2G)   3 PC (CPU: 1 , RAM:2G)

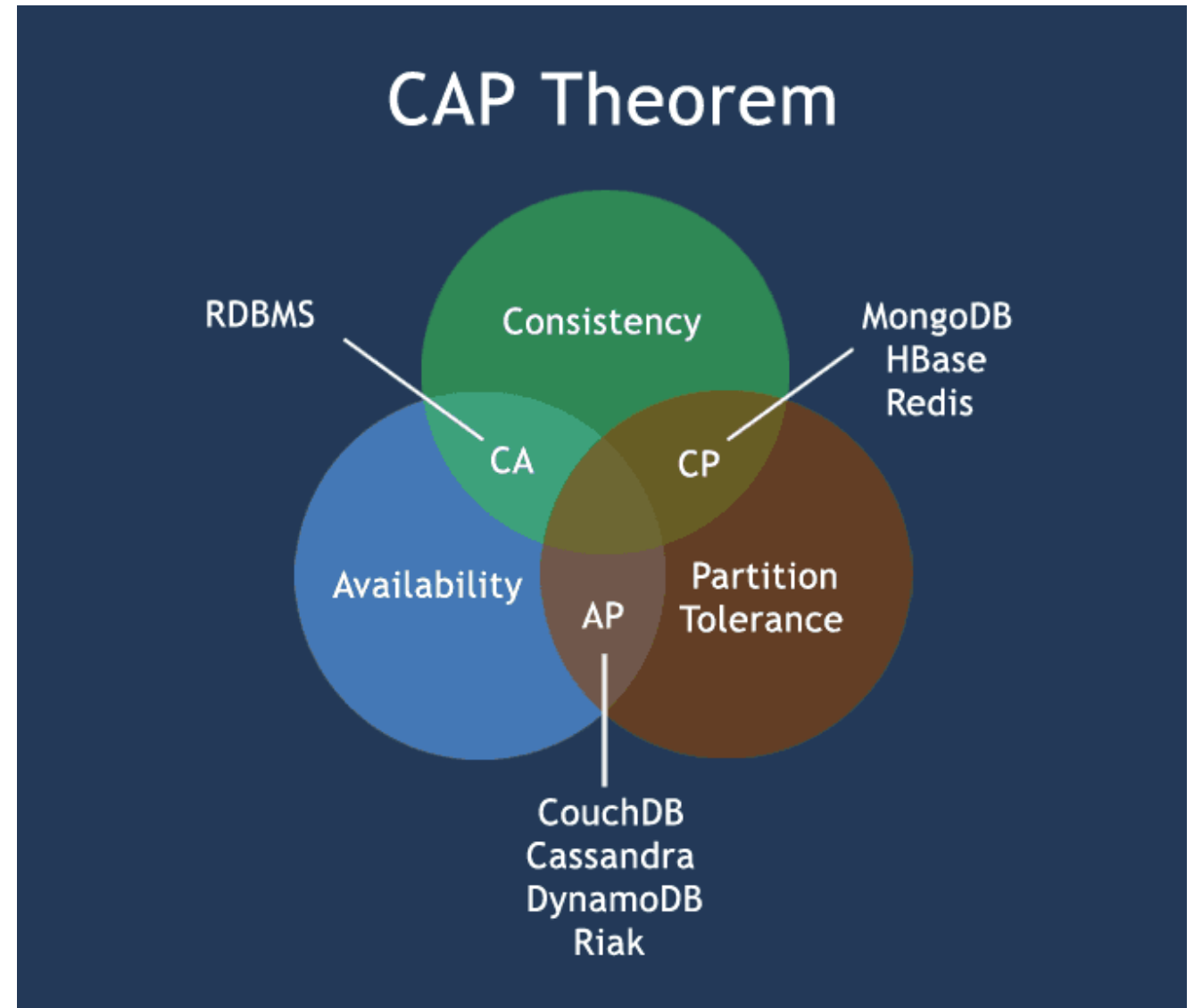**Horizontal Scaling refers to adding additional nodes or machines to your infrastructure to cope with new demands.**

✓ adding a new computer to a distributed software application

NoSQL Databases

@2006-7

- **Stands for Not Only SQL** (beyond sql)

- **Manages structured and semi-structured data.**

- **Prioritizes high performance, high availability and scalability**

- **Designed for Horizontal scaling. Reliable, fault tolerant, Better performance/Speed.**

- **No declarative query language**

- **Uses: Huge data (TBs), Many Read/Write ops, Scalable, Flexible schema.**

- **Don't use if: Need high consistency, Multiple relations**

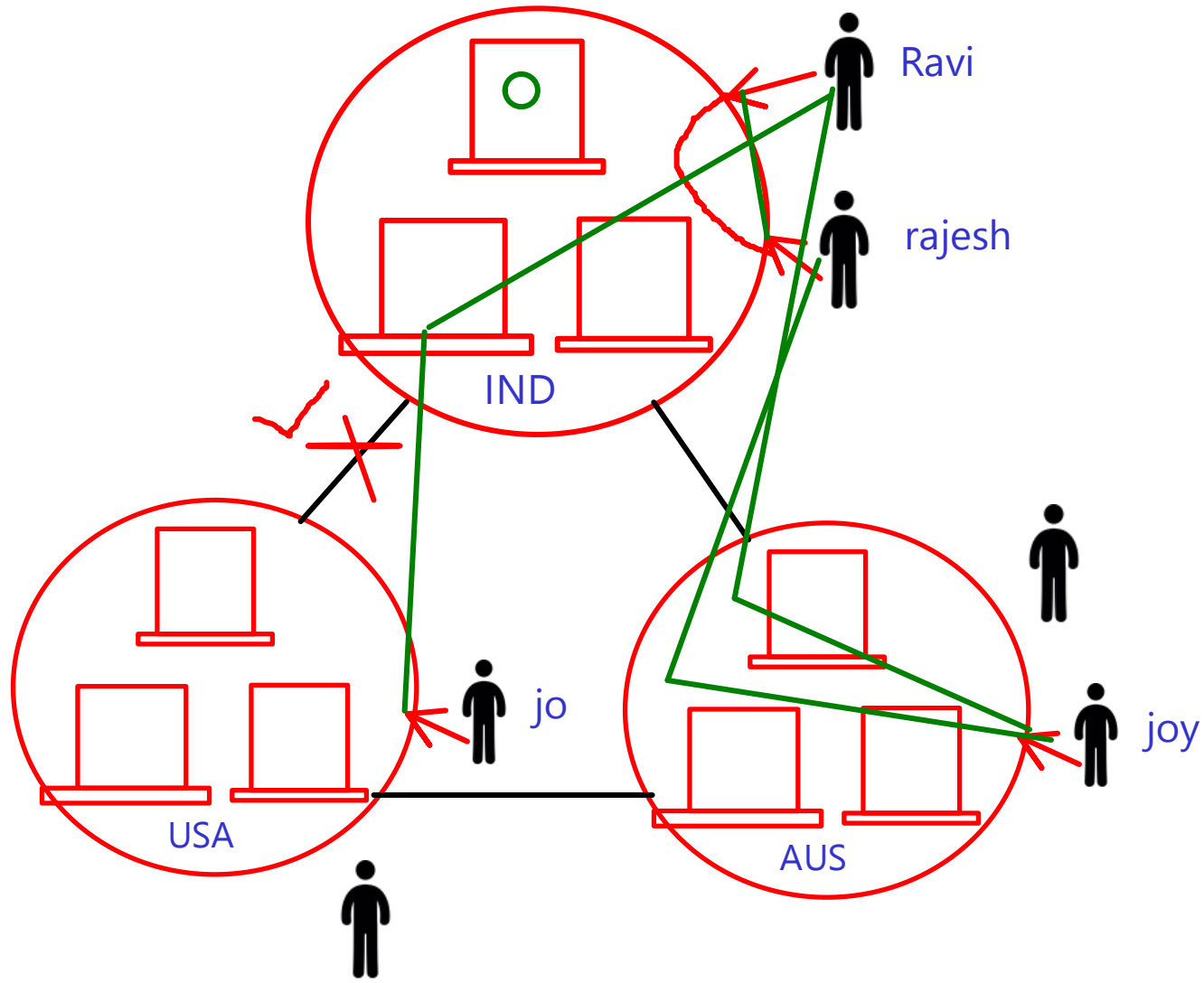- **BASE transactions and Based on CAP Theorem**

CAP Theorem   (brewer's th)

- **C**onsistency - Data is consistent after operation. After an update operation, all clients see the same data.

- **A**vailability - System is always on (i.e. service guarantee), no downtime.

- **P**artition Tolerance - System continues to function even the communication among the servers is unreliable.

CAP Theorem

RDBMS — CA

Consistency

MongoDB
HBase
Redis

CP

Availability

AP

Partition Tolerance

CouchDB
Cassandra
DynamoDB
Riak

FaceBook

BASE

BA = Basically Available

  system running 24 X 7

S => Soft state
  Data is auto transferrd to all nodes in cluster

E => Eventual consistency

  same data visible to all cilent Eventually

MCQ

Q: 1. ROLLBACK is _____ type command .

A. DCL

B. TCL

C. DDL

D. DDD

MCQ

Q: 1. ROLLBACK is _____ type command .

A. DCL

B. TCL

C. DDL

D. DDD

MCQ

Q: 1. _____ command is used to delete table.

A. FREE

B. DELEET

C. DROP

D. RELISED

MCQ

Q: 1. _____ command is used to delete table.

A. FREE

B. DELEET

C. DROP

D. RELISED

Q: 1. Which one is not in v's of Big Data?

A. Variety

B. Velocity

C. Volatile

D. volume

Q: 1. Which one is not in v's of Big Data?

A. Variety -> Data can be unstructured, semi-structured or structured

B. Velocity -> Data generated with high speed

C. Volatile

D. volume -> Hugh amount of data

Veracity in Big data means_____.

A. The data is generated with high speed

B. The data is huge

C. The data is reliable and trustworthy

D. The data management

Veracity in Big data means_____.

    A. The data is generated with high speed -> Velocity

    B. The data is huge -> volume

    C. The data is reliable and trustworthy -> Veracity

    D. The data management -> database

In RDBMS data is stored in _____.

A. document

B. tables

C. collection

D. keys

In RDBMS data is stored in _____.

A. document

B. tables

C. collection

D. keys

_____ introduced the NoSQL concept in 1998.

A. Cassandra

B. Devid Sam

C. Carlo Strozzi

D. E.F. CODD

_____ introduced the NoSQL concept in 1998.

A. Cassandra

B. Devid Sam

C. Carlo Strozzi

D. E.F. CODD

Thank you!

Pradnyaa S. Dindorkar   <pradnya@sunbeaminfo.com>