



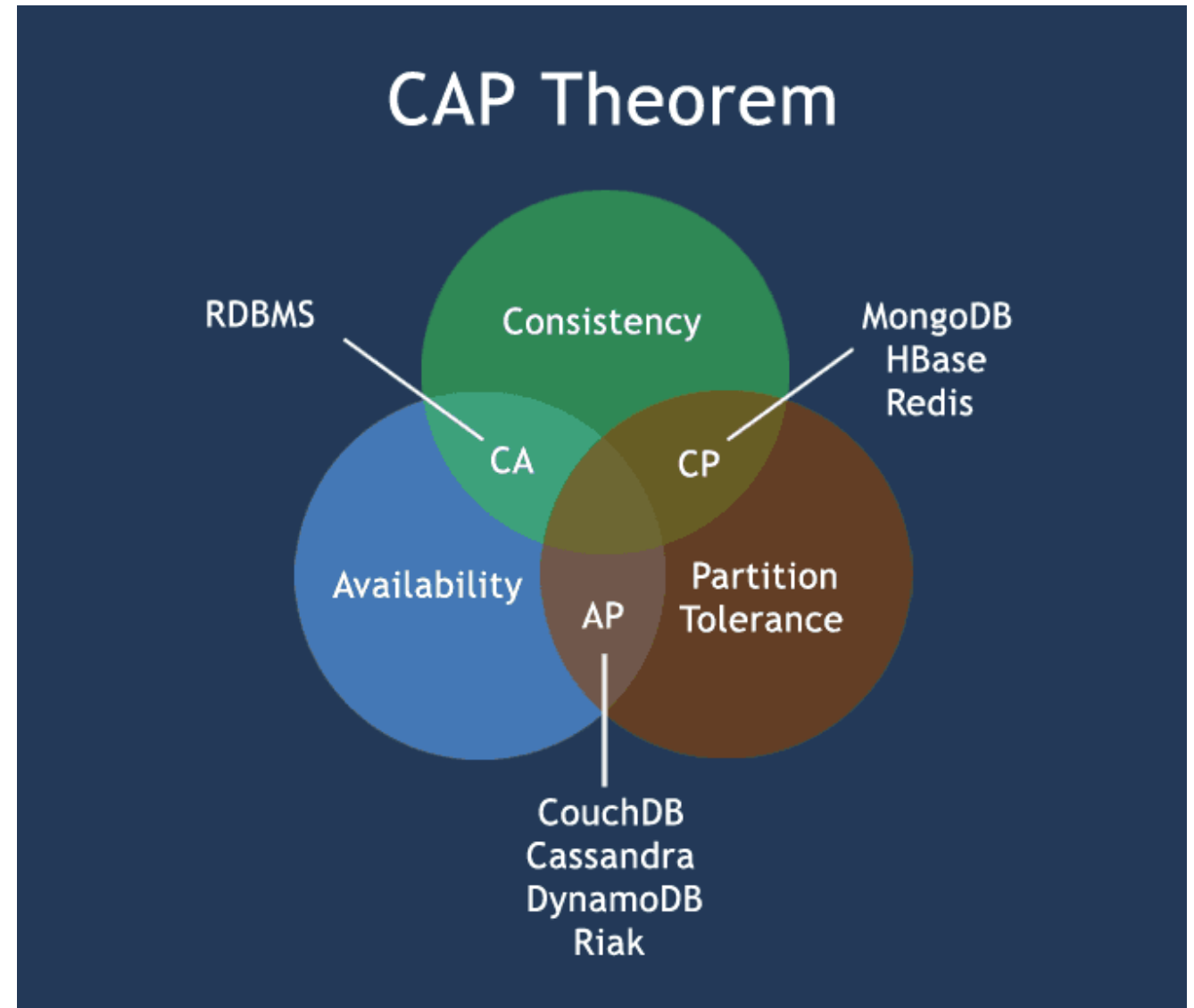
# **Fundamentals of Data Engineering**

**Trainer: Pradnyaa S Dindorkar**



# CAP Theorem

- **Consistency** - Data is consistent after operation. After an update operation, all clients see the same data.
- **Availability** - System is always on (i.e. service guarantee), no downtime.
- **Partition Tolerance** - System continues to function even the communication among the servers is unreliable.



# NoSQL Databases

- Key-value databases - e.g. redis, dynamodb, riak
  - Based on Amazon's Dynamo database.
  - Keys are unique and values can be of any type i.e. JSON, BLOB, etc.
  - Implemented as big distributed hash-table for fast searching.

Product ID	Type	Schema is defined per item		
1	Book ID	Odyssey	Homer	1871
2	Album ID	6 Partitas	Bach	
2	Album ID: Track ID	Partita No. 1		
3	Movie ID	The Kid	Drama, Comedy	Chaplin

50000 X 30      pune=1000  
RDBMS=50000 + (30)

- Wide Column databases - e.g. hbase, cassandra, bigtable, ...
  - Values of columns are stored contiguously.
  - Better performance while accessing few columns & aggregations. *count, sum, min, max, avg*
  - Good for data-warehousing, business intelligence, CRM, ...

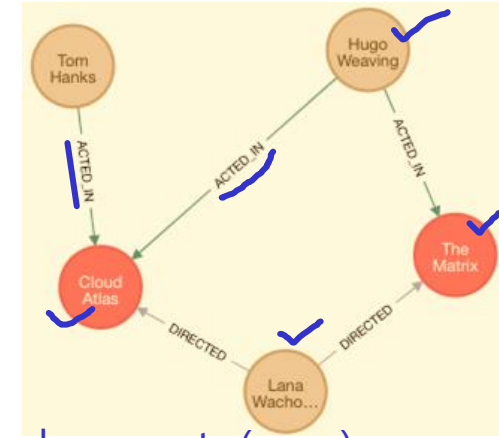
row ID

Row Key	Customer		Sales	
Customer Id	Name	City	Product	Amount
101	John White	Los Angeles, CA	Chairs	\$400.00
102	Jane Brown	Atlanta, GA	Lamps	\$200.00
103	Bill Green	Pittsburgh, PA	Desk	\$500.00
104	Jack Black	St. Louis, MO	Bed	\$1600.00

# NoSQL Databases

- Graph databases - e.g. Neo4J, Titan, ...

- Graph is collection of vertices and edges.
- Excellent performance, while dealing with all relations of an entity  
(irrespective of size of data).

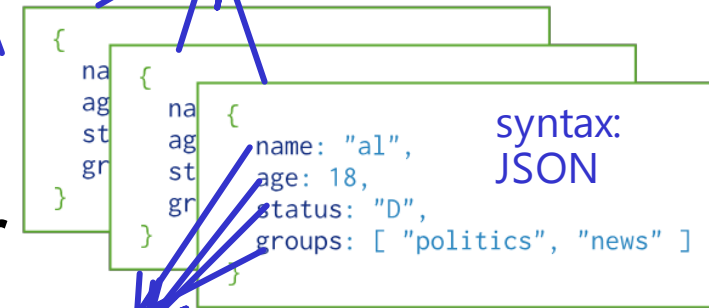


- Document oriented databases - e.g. MongoDb, CouchDb, ...

- Document contains data as key-value pair as JSON or XML.
- Document schema is flexible & are added in collection for processing.

Javascript Object Notation

semi-structured



documents (rows)

fields

Collection (table)

(columns)



# Document – flexible schema

mobile m1

```
{
  CPU: -----,
  RAM:-----,
  Keypad:----,
  Display:----,
}
```

mobile m2

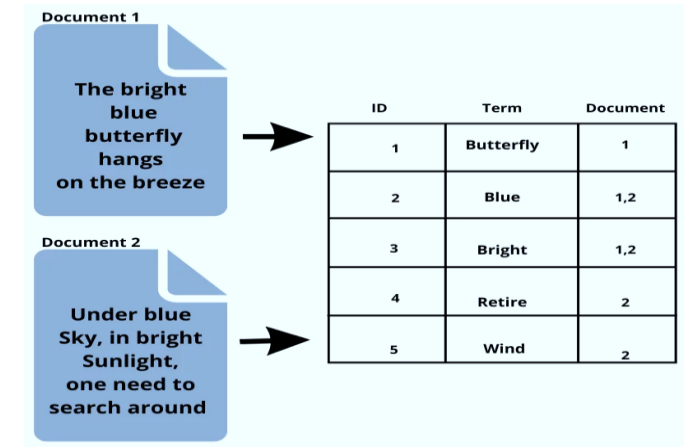
```
{
  CPU: -----,
  RAM:-----,
  Display:----,
  Bth:-----,
  GPS:-----,
  OS:-----,
  Camera:-----,
}
```

RDBMS	DOC-NoSQL
Table	collection
rows	documents
columns	fields



# NoSQL Databases

- ✓ **Search databases – e.g. Elasticsearch, Solr, Lucene, ...**
  - For faster search – Text search, Log analysis.
  - Indexed, Exact/Fuzzy matches, Anomaly detection, Analytics.
- ✓ **Time series databases – e.g. Influx, Druid, ...**
  - Values organized by time like stock market, weather, ...
  - Optimized for retrieval, statistical processing, ...
  - Used for measurement data (weather, ...) and event-based data (accidents, ...)

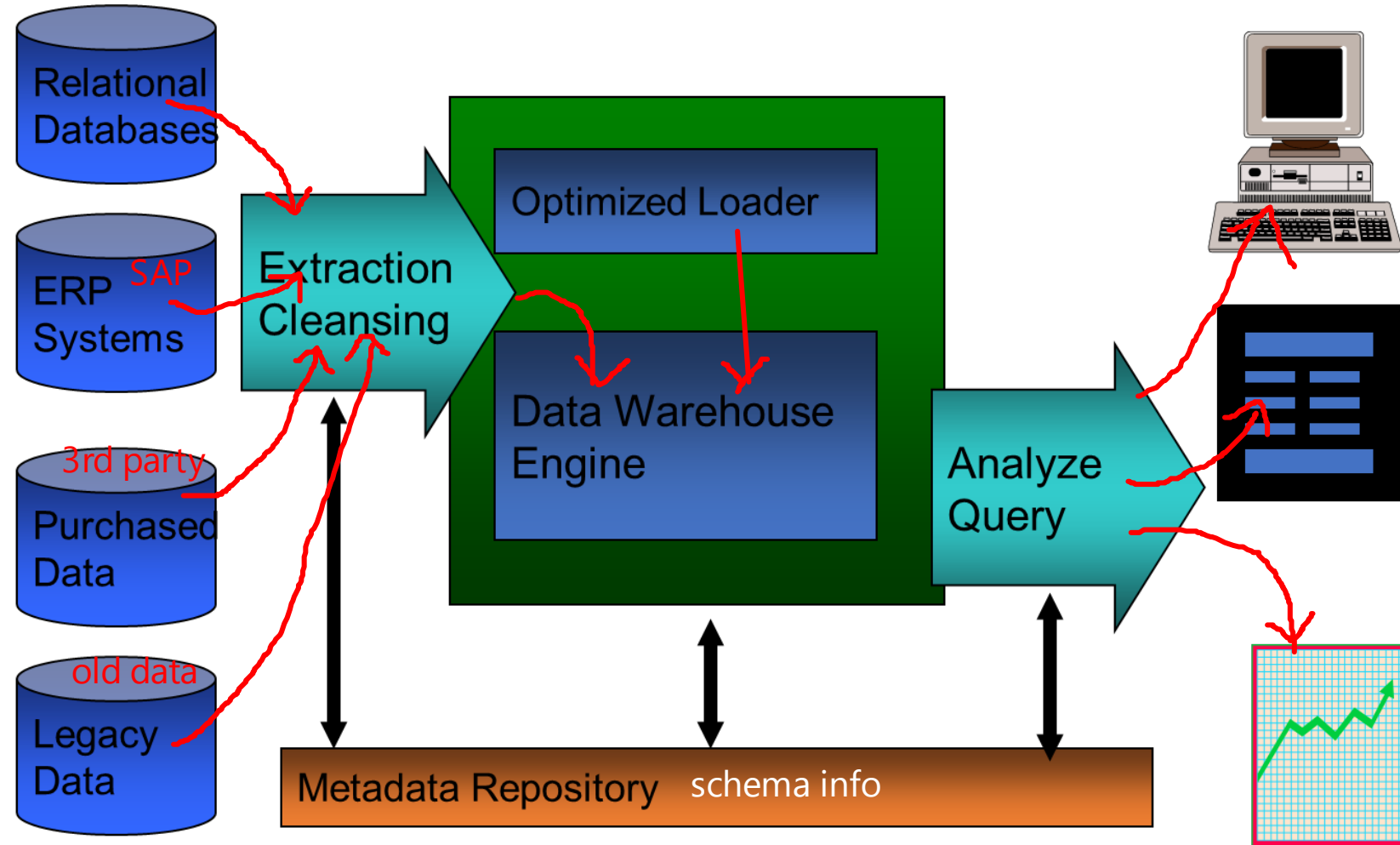


Date	Ozone (µg/m³)	Temperature (°C)	Relative humidity (%)	n deaths
1 Jan 2002	4.59	−0.2	75.7	199
2 Jan 2002	4.88	0.1	77.5	231
3 Jan 2002	4.71	0.9	81.3	210
4 Jan 2002	4.14	0.5	85.4	203
5 Jan 2002	2.01	4.3	93.5	224
6 Jan 2002	2.4	7.1	96.4	198



# Data warehousing

- Data warehouse is a single, complete and consistent store of data obtained from a variety of different sources made available to end users in a what they can understand and use in a business context.
- Data warehousing is a process of transforming data into information and making it available to users in a timely enough manner to make a difference.



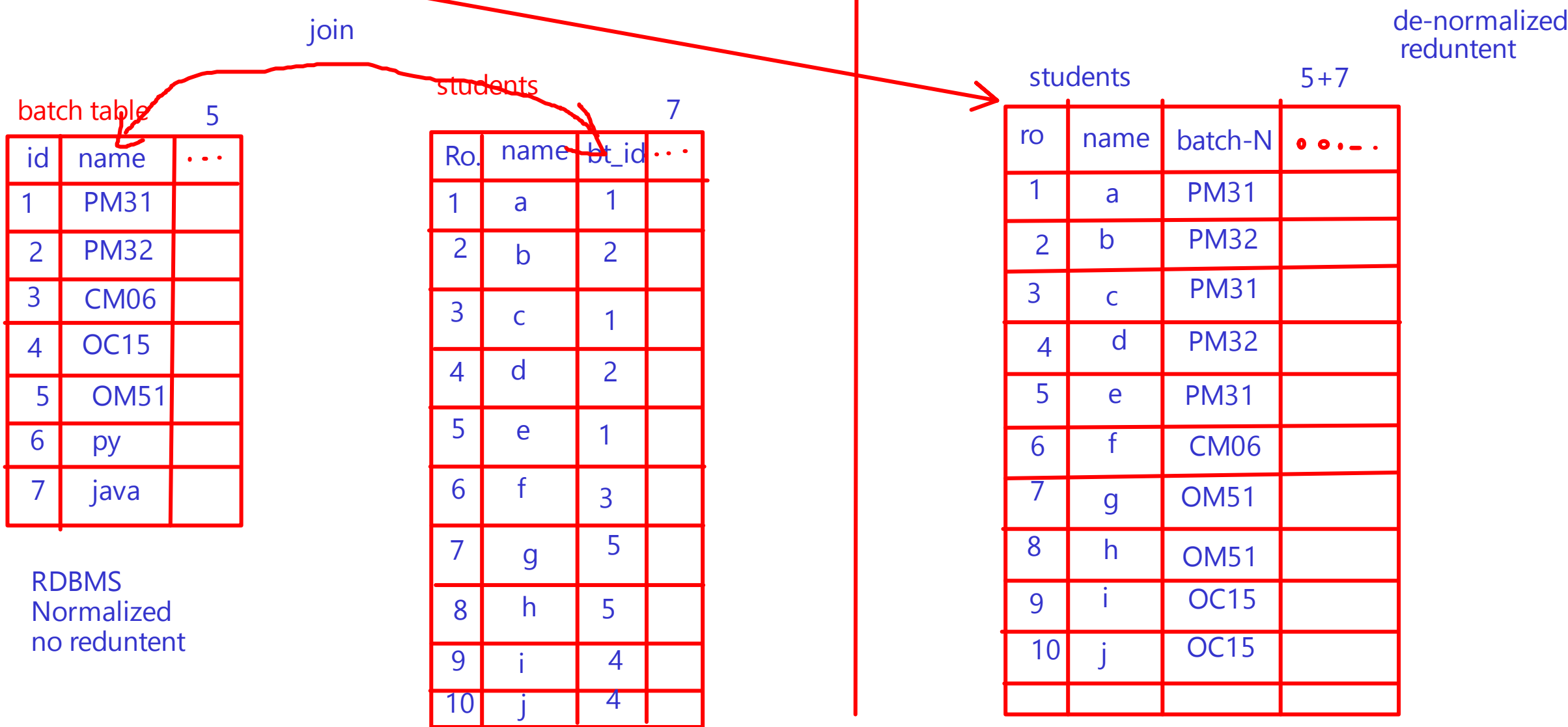
# Extract – Transform – Load

- ✓ Extracting: Extract data from sources into staging area temp memory
- ✓ Conditioning: Data types conversion to fit warehouse.
- ✓ House holding: Grouping similar data 1000 / 4 = 250 1/4
- ✓ Enrichment: Add relevant data from external sources
- ✓ Scoring: Computation of probability of an event
- ✓ Scrubbing: Data cleaning: find duplicate, missing data
- ✓ Merging: Merging data from various sources.
- ✓ De-normalize: Duplicate data to reduce joins.
- Loading: Load data in warehouse models like Star, Snowflake, Galaxy.
- ✓ Delta Updating: Incremental data uploading
- ✓ Partitioning: Dividing the data in logical parts to improve performance.





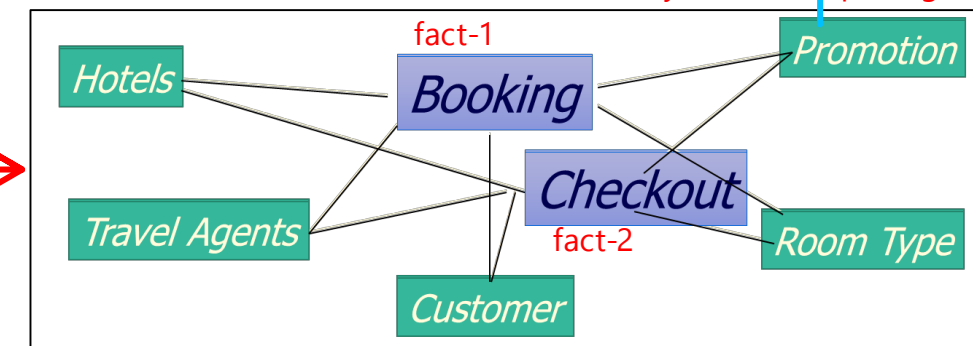
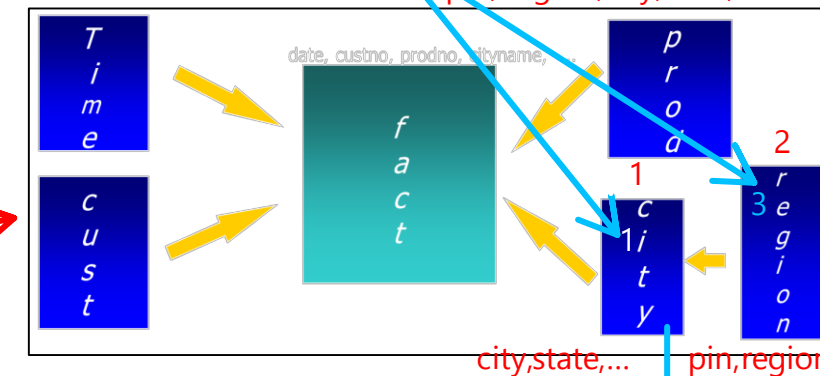
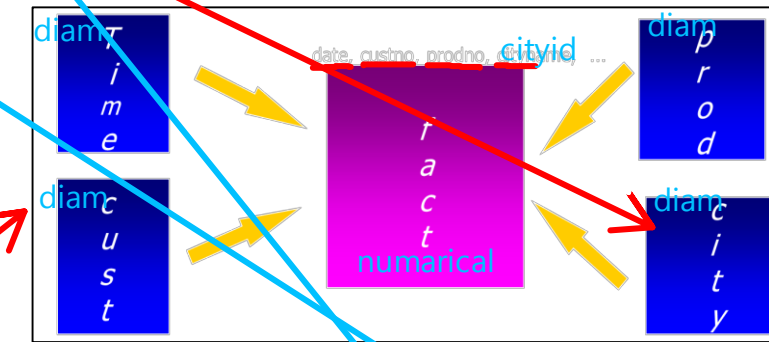
# De-normalize :- Duplicate data to reduce joins.



# DWH Schemas

411065	hinjewadi	pune	MH	IND
411055	market yard	pune	MH	IND
411000	katraj	pune	MH	IND

- DWH schema is how data is stored in tables in warehouse for the efficient processing of the data.
- A fact table stores metrics, measurements, or facts about business processes. *only numerical*
- Dimension tables are tables used to store data attributes or dimensions. *numerical + string*
- Star schema: Single facts table and a few dimension tables (de-normalized) – Simple design.
- Snowflake schema: Single facts table and connected dimension/sub-dimension tables (normalized).
- Galaxy or Fact-Constellation schema: Multiple facts tables mapped to multiple dimension/sub-dimension tables.



# OLTP (Database) vs OLAP (Data warehouse)

- Online Transaction Processing

- Modeled to run the business
- Detailed/Transactional normalized real-time data
- Transaction performance DML  
DQL
- Read/Write operations
- Isolated data (Application specific)
  - Limited data (100 MB to 100 GB)

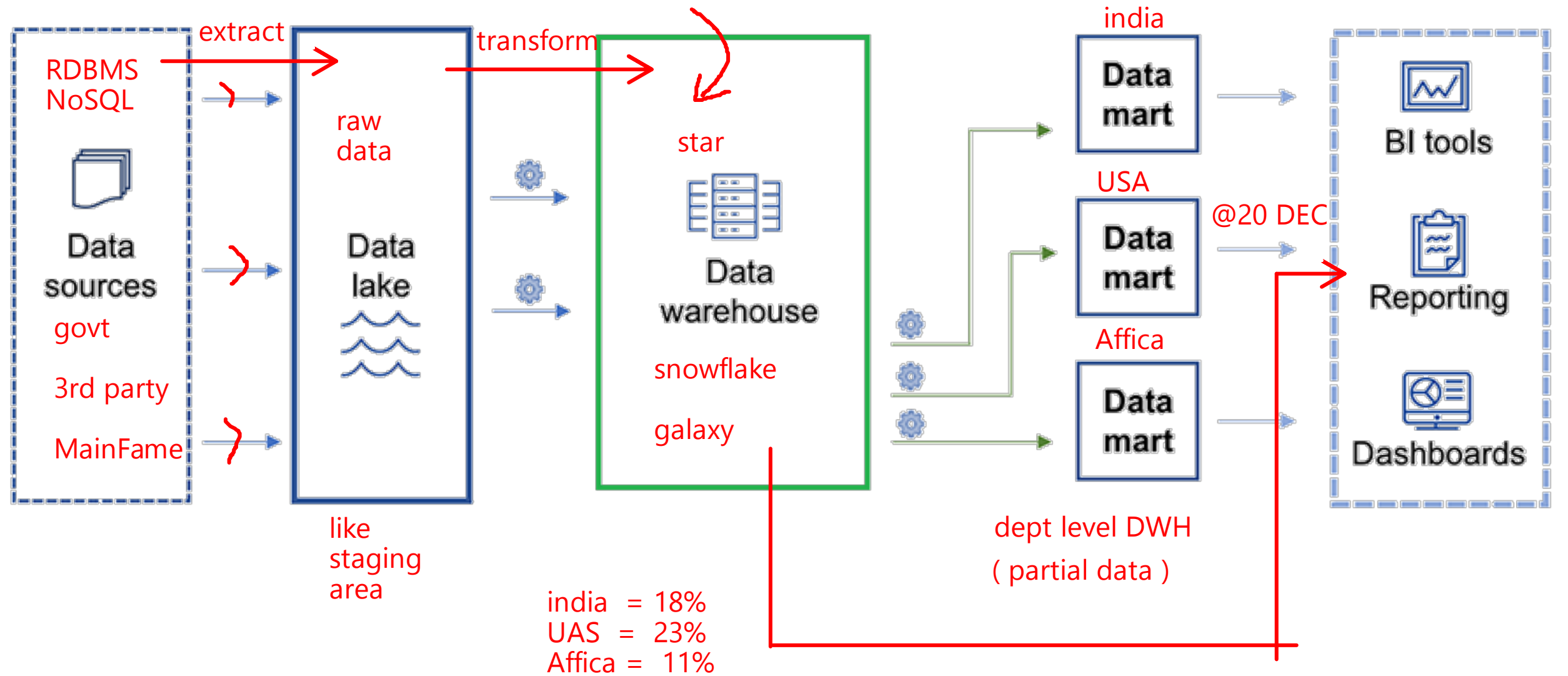
- Online Analytical Processing

- Modeled to analyze/optimize business
- Summarized/refined redundant snapshot data historical data
- Analytical query performance DQL  
joins  
aggr
- Mostly Read operations sum,avg
- Integrated data (from all sources) – Huge data (100 GB to Few TB)



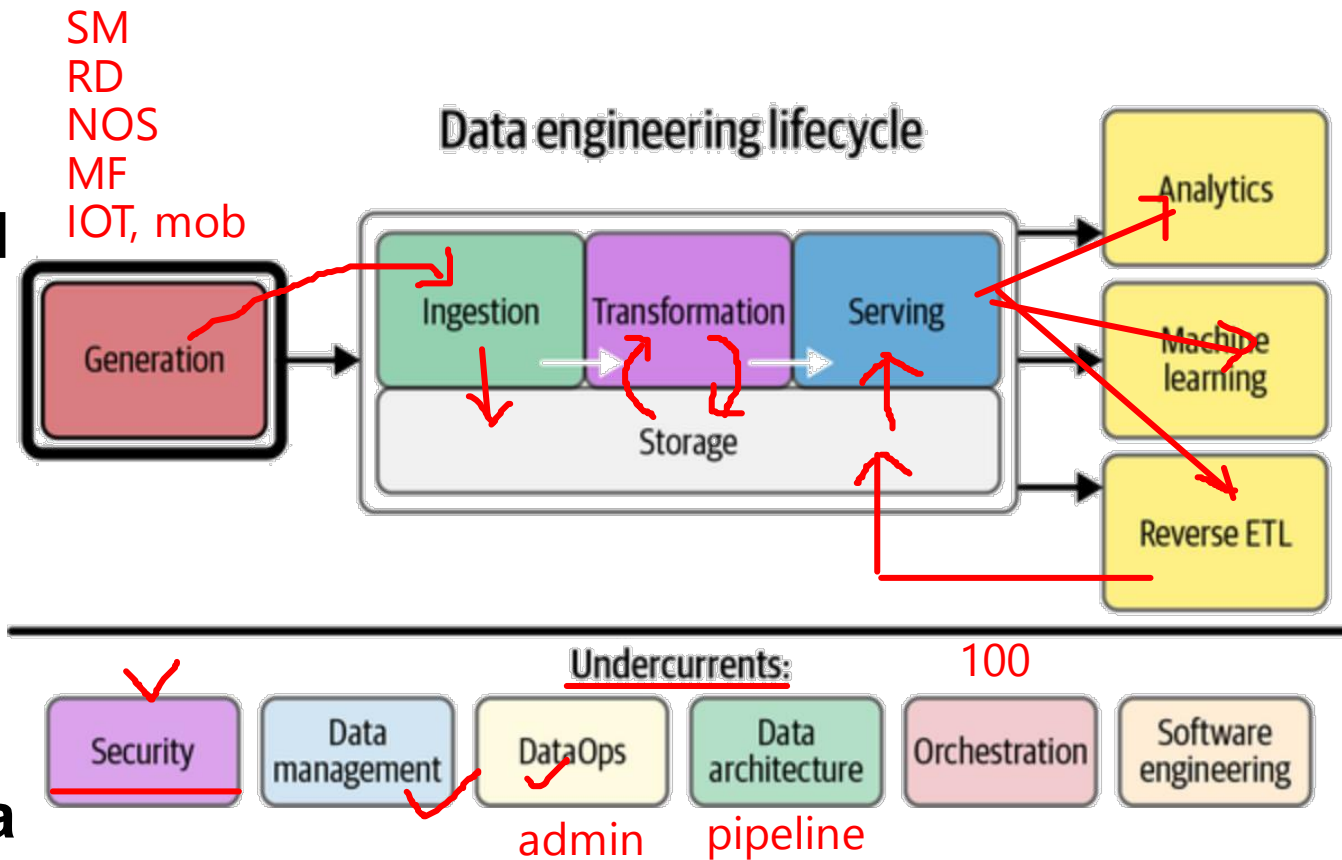
# Data lake vs Data warehouse vs Data mart

@ 20 Oct



# Data engineering

- Data engineering is the development, implementation, and maintenance of systems and processes that take in raw data and produce high-quality, consistent information that supports downstream use cases, such as analysis and machine learning.
- Data engineer manages data engineering lifecycle, beginning with getting data from source systems & ending with serving data for use cases, such as analysis or machine learning.



# Traditional ETL vs Hadoop ELT

- ETL stands for Extract, Transform and Load.
- The ETL process typically extracts data from the source/transactional systems, transforms it to fit the model of data-warehouse and finally loads it to the data warehouse.
- The transformation process involves cleansing, enriching and applying transformations to create desired output.
- Data is usually dumped to a staging area after extraction.
- ELT stands for Extract, Load and Transform.
- As opposed to loading just the transformed data in the target systems, the ELT process loads the entire data into the data lake. This results in faster load times.
- The load process can also perform some basic validations and data cleansing rules.
- The data is then transformed for analytical reporting as per demand.



# Data storage

- Data storage is related to multiple stages in data engineering life cycle i.e. ingestion, transformation and serving.
- Storage needs to be selected based on read/write requirement, speed, durability, consistency, availability, scalability, fault tolerance, ... factors.
- Storage tradeoffs
  - Local storage vs Distributed storage
  - Strong consistency vs Eventual consistency
- Storage options are: File storage, Local disk storage, Network attached storage (NAS), Cloud file systems (S3/Blob), Block storage, RAID, Storage area network (SAN), Object storage, HDFS, Streaming storage.



---

Q: Which of the following is not the type of NoSQL?

A: Graph

B: Document

C: Text

D: Search





Q: Which of the following is not the type of NoSQL?

A: Graph

B: Document

C: Text

D: Search



---

Q: Which of the following is the type of NoSQL?

A: Row

B: Column

C: Collection

D: Table



---

Q: Which of the following is the type of NoSQL?

A: Row

**B: Column**

C: Collection

D: Table



---

Q: Which of the following is not the valid Data warehouse schema?

A: Star

B: Showfall

C: Showflake

D: Galaxy



Q: Which of the following is not the valid Data warehouse schema?

A: Star

**B: Showfall**

C: Showflake

D: Galaxy



Q: Which of the following is working on real time data?

A: OLAP

B: OLTP

C: OLPP

D: OTLP



Q: Which of the following is working on real time data?

A: OLAP

**B: OLTP**

C: OLPP

D: OTLP



Q: Analytical queries are performed in \_\_\_\_\_.

A: OLAP

B: OLTP

C: OLPP

D: OTLP





Q: Analytical queries are performed in \_\_\_\_\_.

A: OLAP

B: OLTP

C: OLPP

D: OTLP



Q: In data engg life cycle pulling data is called known as \_\_\_\_\_ data.

A: Popping

B: Ingesting

C: Serving

D: Analysing



Q: In data engg life cycle pulling data is called known as \_\_\_\_\_ data.

A: Popping

**B: Ingesting**

C: Serving

D: Analysing





**Thank you!**

**Pradnyaa S Dindorkar <[pradnya@sunbeaminfo.com](mailto:pradnya@sunbeaminfo.com)>**

