## Agenda

a. Parquet Files demo in hive
b. Data warehousing in cloud Redshift
c. Distributed system and core components in Hadoop.
d. YARN architecture in detail

# Parquet File Format

→ AVRO → JSON File

→ text
→ Binary
→ Avro

→ Parquet file Format

→ Columnar Format

| id | name | Sal | Dept | Locaha |
|---|---|---|---|---|
| 1 | Amit | 7000 | IT | Delhi |
| 2 | Amit R | 9000 | IT | Ahemdabad |
| 3 | Rakesh | 10000 | IT | Kochi |
| 4 | Abhishek | 10000 | IT | Panaji |
| 5 | Ayush | 9000 | AD | Delhi |

Ayush → location, Sal, Dept

100,000

Delhi, 9000, AD

25 cols

25 cols

1, 2, 3, 4, 5

Amit, AmitR, Rakesh, Abhishek, Ayush, Rita

7000, 9000, 11000, 10000, 5000, 11000

id

name

Sal

loc

5 cols

5, Ayush, –, Delhi

Delhi

5000

Null

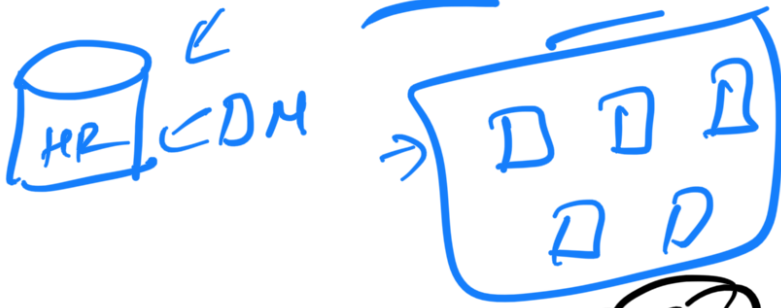
CRUD ⇒ Update

User profile  Avro

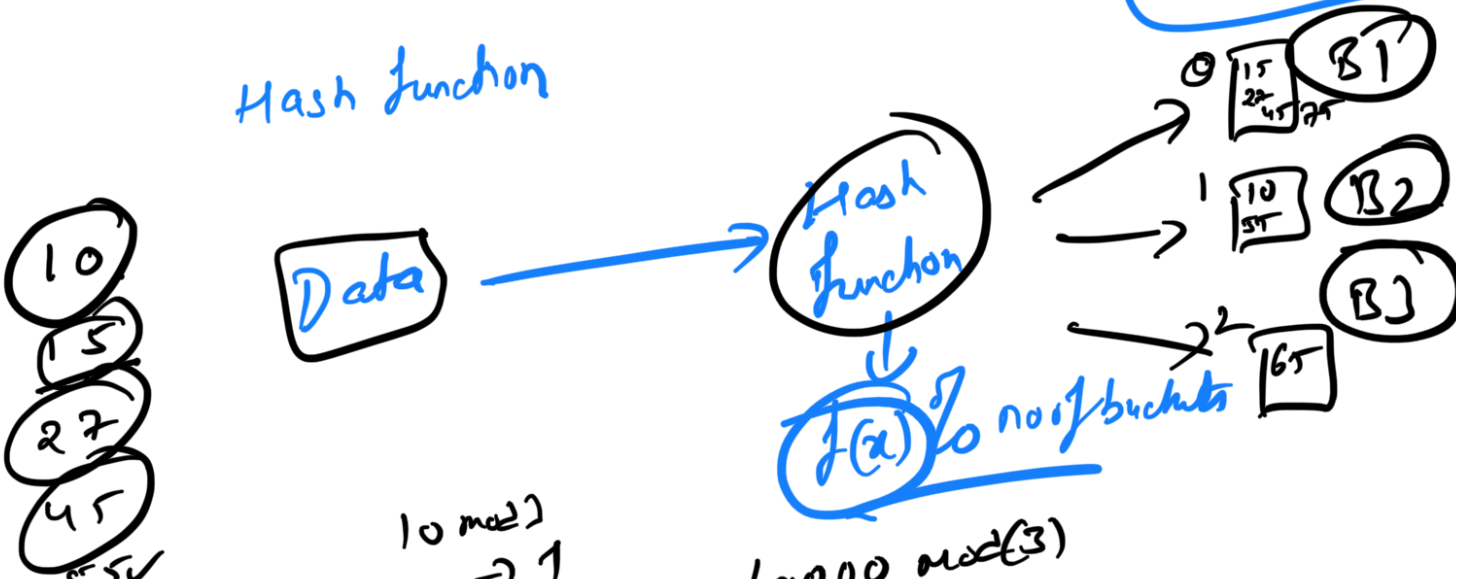
Registration_dttm = Acr ---UA | AF4wUA
id = 1 | 2

④ Bucketing :-


Rakesh →




Hash function



65

(5) Cost Based optimizer

(1) Query Vectorization

Set hive.vectorized.execution.enabled = true; (false)

→ one by one
↳ Batch

Hive
↳ Tiktok
↳ Walmart → $1 billion

CDH → Apache Hive

AWS = ~~Redshift~~

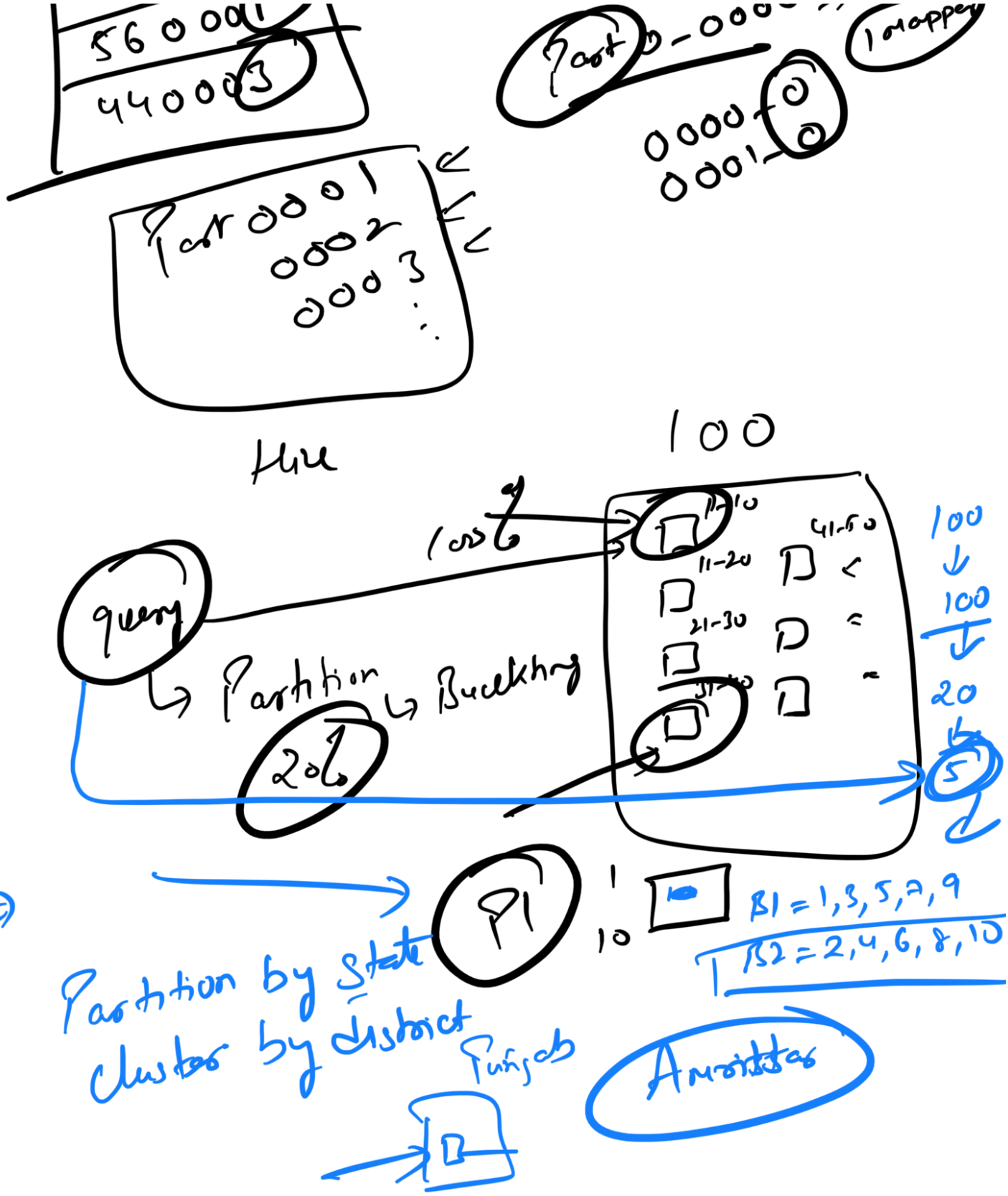
Azure = Synapse

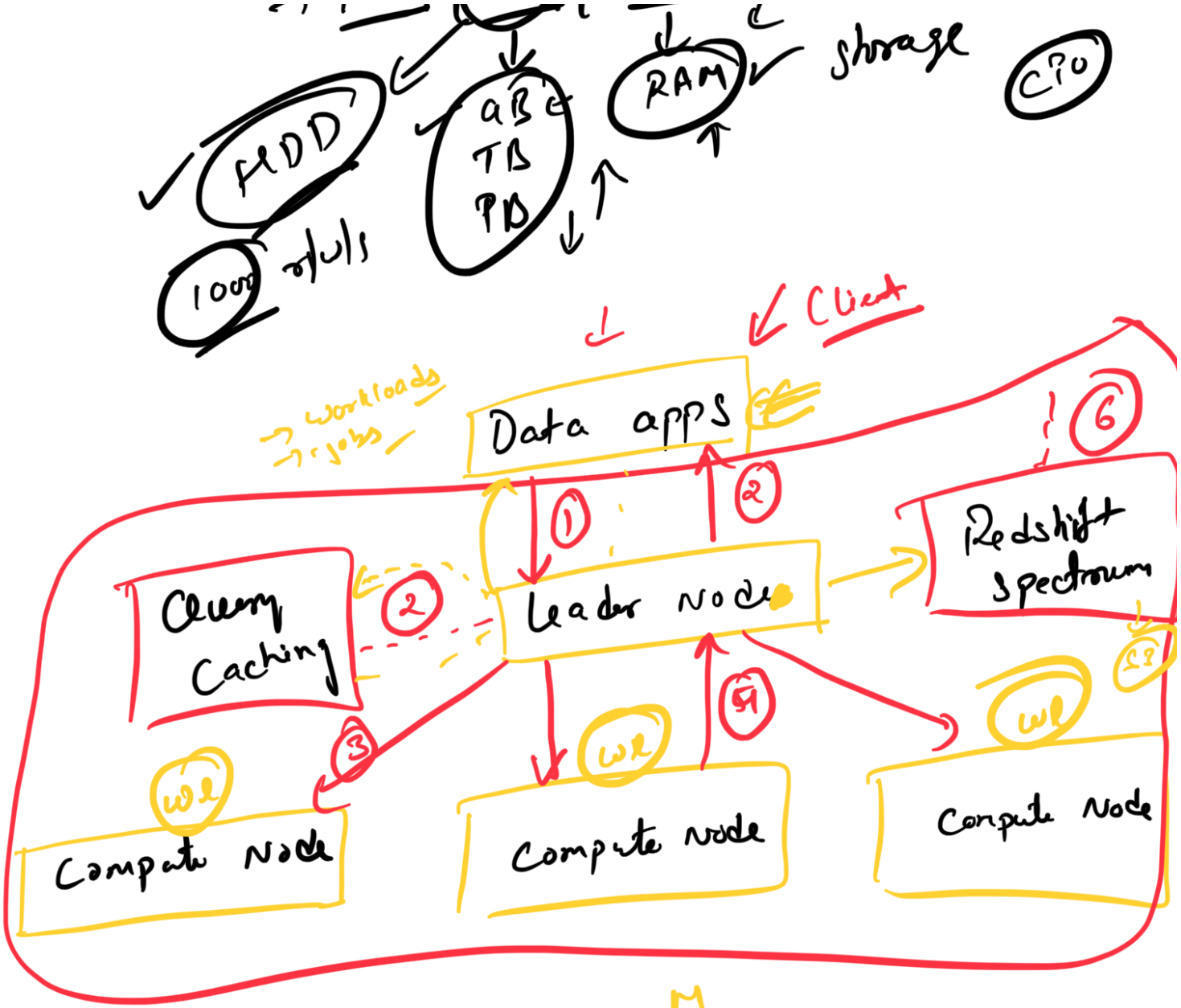
GCP = Big Query

} AWS I SQL

%10 ⇒ 8 → 0008-0

121008
122005


100
query
Partition
Bucketing
20
100
100
20
Partition by State
Cluster by district
Punjab
Amritsar
Aws Redshift
Hardware
SSD
6000



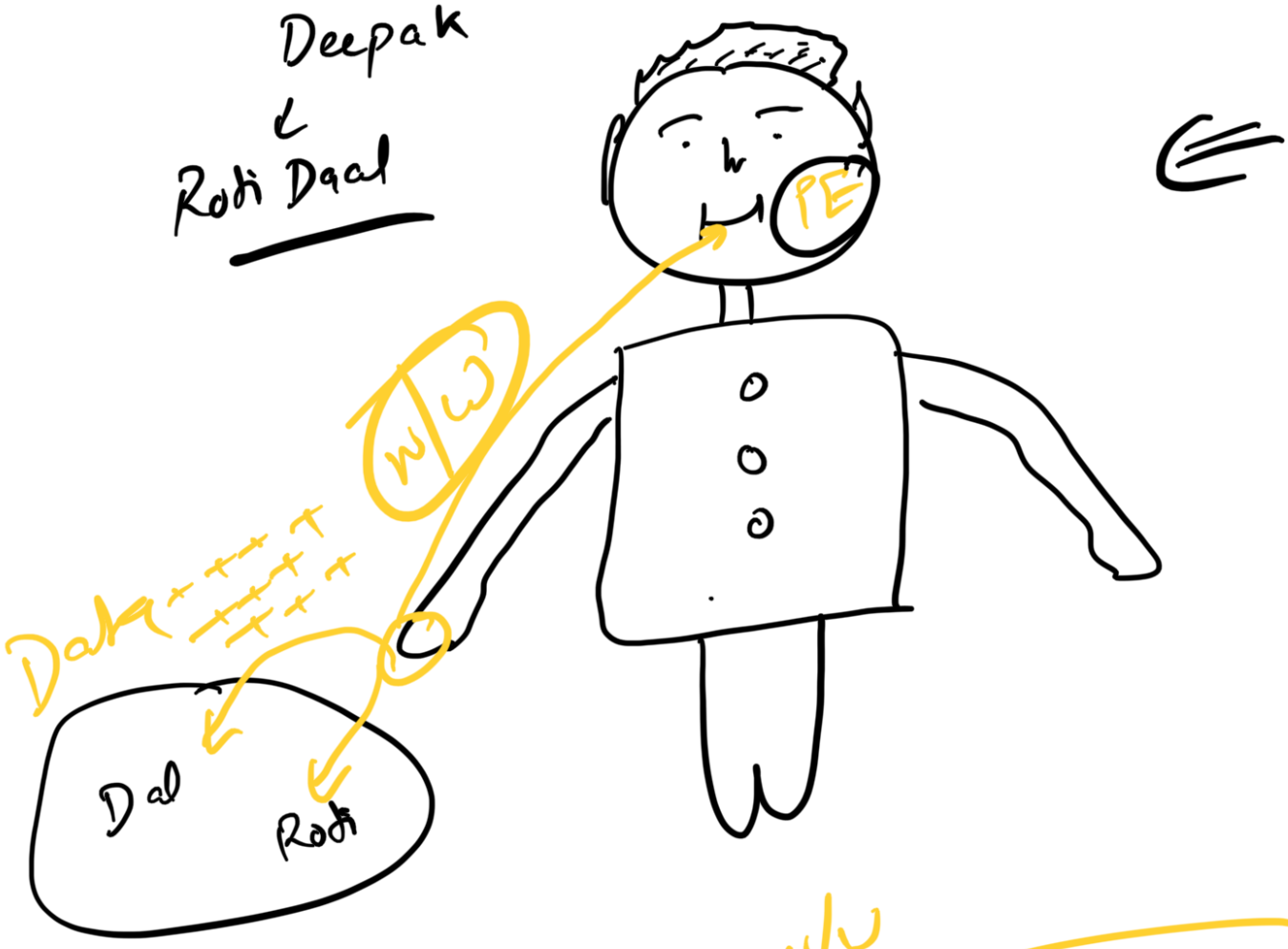

① Leader Node → M

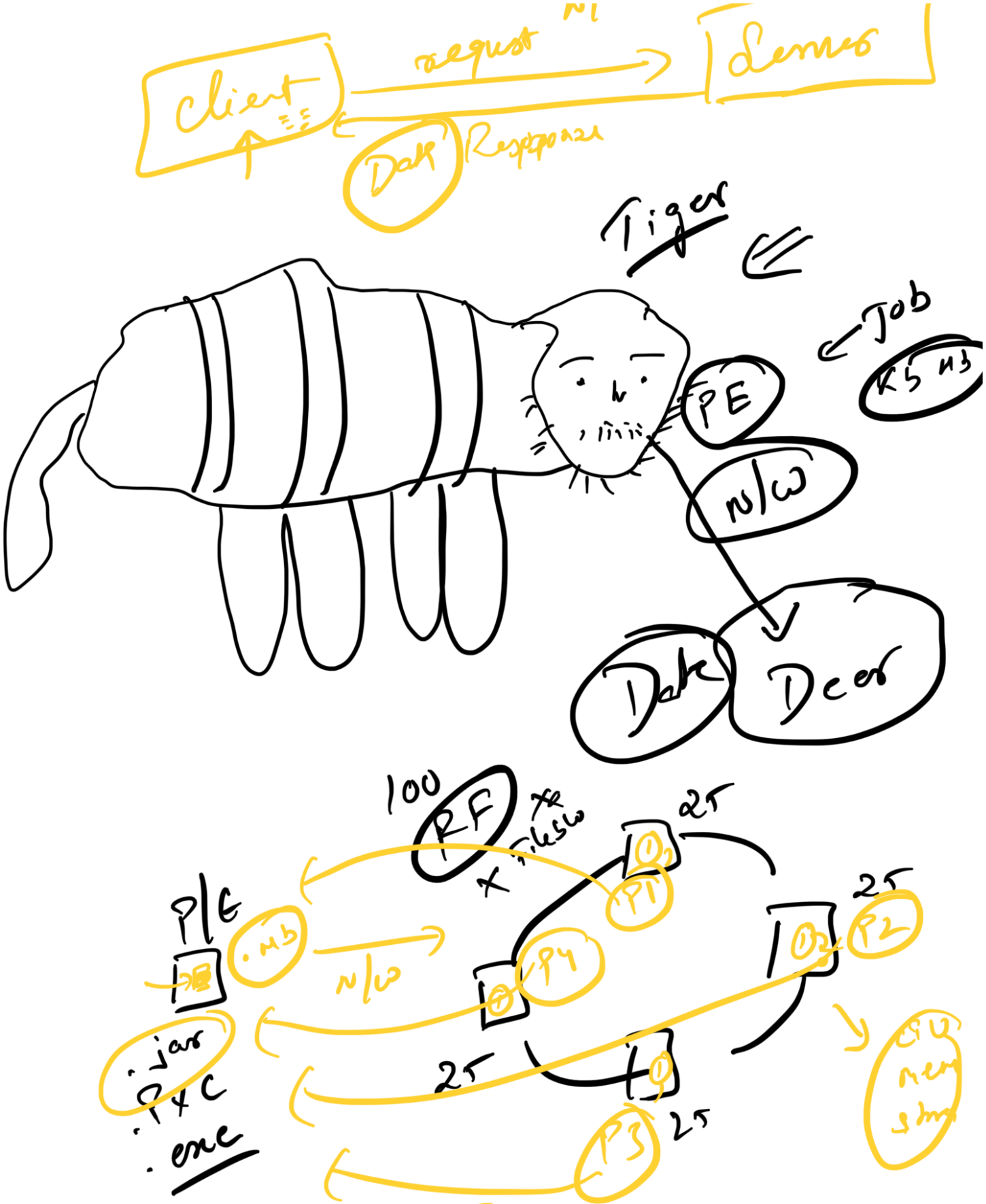
② Compute Node → S




Redshift Spectrum

## How Hadoop Process Data ?




Client
request
Server
Data
Response
Tiger
Job
PE
N/W
Data
Deer
100
RF
P1
P4
P2
P3
25
25
25
25
P/C
.jar
.P4C
.exe
N/W
MB
Hadoop → Ecosystem

16-18
YARN
MR
HDFS
Core
Spark