Agenda

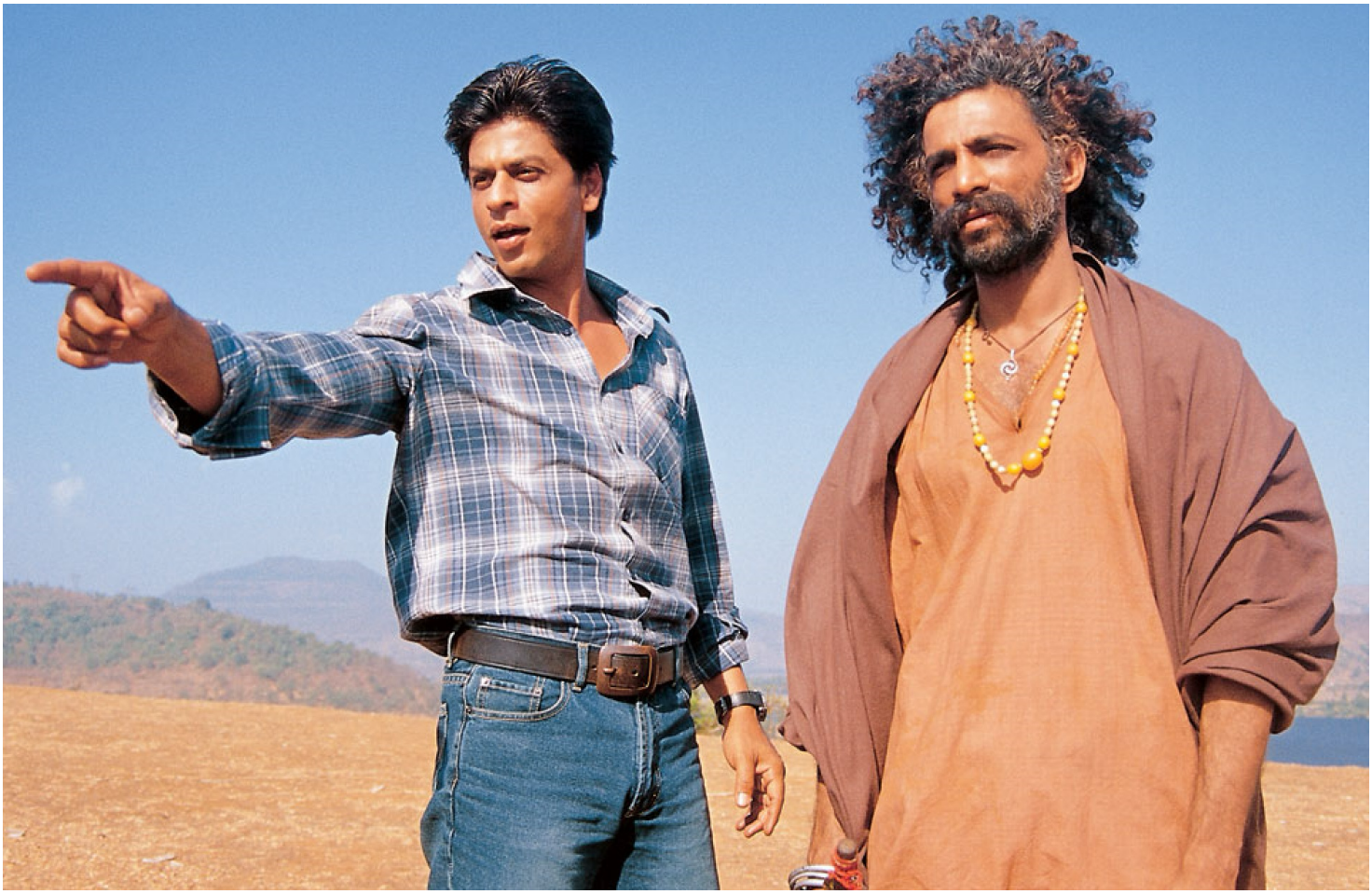
a. Optimisations in Hive

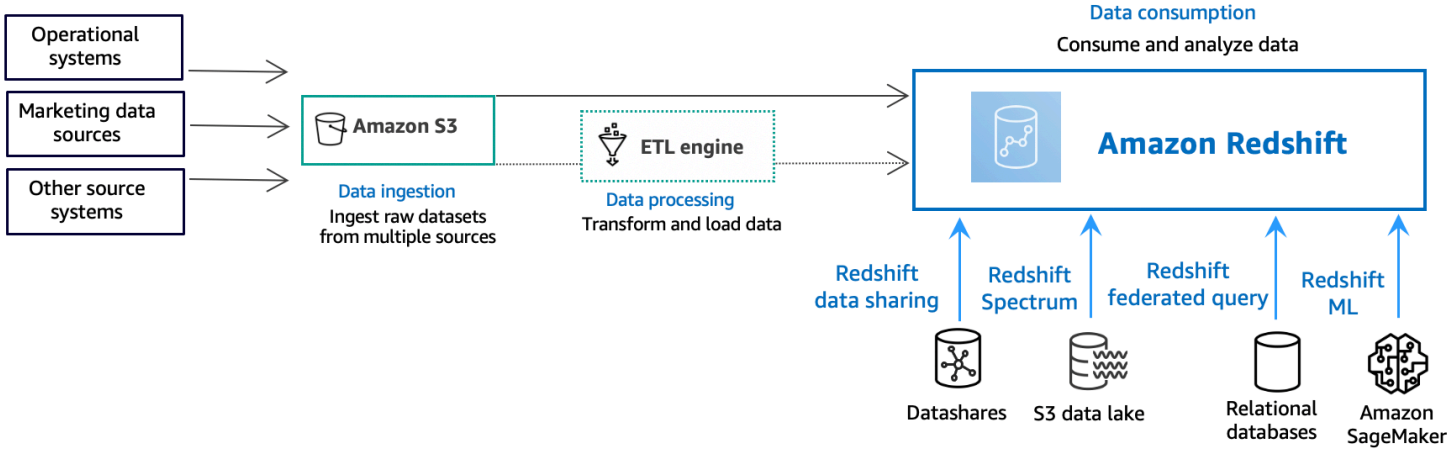
b. **Essential features of Redshift**

C. Columnar Data warehouse concepts

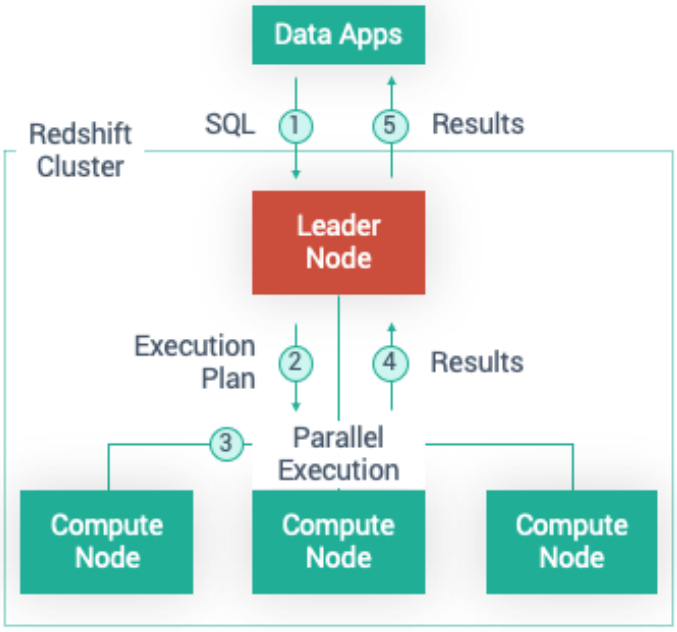
d. Architecture

e. Distribution key and sort key





Amazon Redshift Architecture: The Life of a Query



The 5 Steps to Process a Query

1. The cluster receives a query coming from a data app and parses the SQL in the leader node.
2. The leader node creates an execution plan that breaks a query down into a discrete sequence of steps.
3. The leader node distributes the work of executing the steps in parallel across the compute nodes.
4. The compute nodes send the results back to the leader node to merge data into a single result.
5. The leader node addresses any final sorting or aggregation and returns the results to the data app.

Optimisation

① change the execution Engine
↳ MR / Tez / Spark

② use external managed table

ⓐ Managed table
Hive ⟵ ↳ ✗ Metadata (RDBMS)
↳ ✗ Data (HDFS)

Drop

ⓑ external managed table

Metadata ✓
Hive → ✗ → Data

Drop

③ File types :-

ⓐ text File ✓ ⇒ basic human readable file
↳ CSV, TSV, text
⇒ Consumes more space

ⓑ Sequence File ✓
⇒ key-value but in binary format
⇒ More efficient

ⓒ Avro
ⓓ Parquet

ORC

(1) Avro ⇒ (JSON)
↳ Data Serialization Framework.

Datatypes ⇒ null, boolean, int, long, float, double, bytes, string

Complex Data types

(1) Record (Table) ⇒ A UDT composed of 1 or more named fields.

(2) enum ⇒ Specific set of values

(3) array ⇒

(4) map ⇒

(5) Union ⇒ Exactly one value matching a specified set of types

(6) fixed ⇒ A fixed number of unsigned bytes

A ∪ B = {1, 2, 3, 4, 5}

A = 1, 2, 3

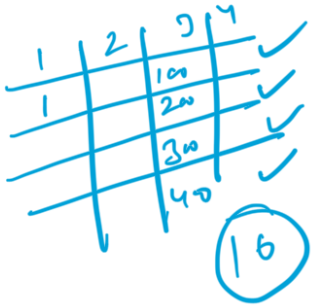
(2) Avro ⇒ (1) = 4, 5, 3

Avro Serializer
Avro Deserializer

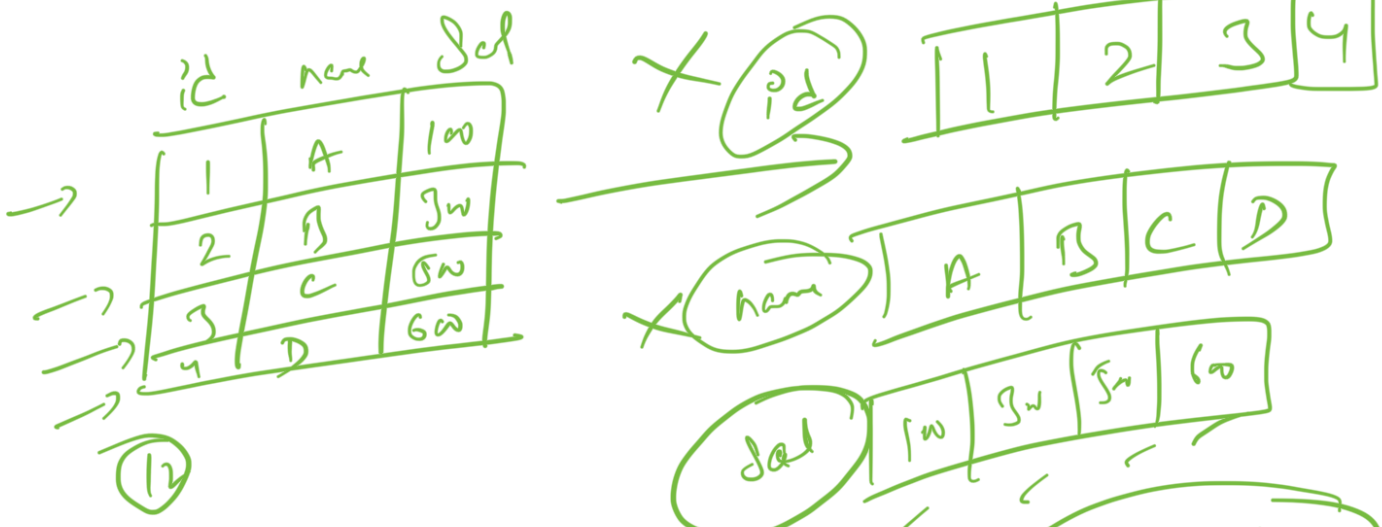
Disadvantages:-

(1) Avro supports row wise storage

(2) To read/write all need schema is needed.



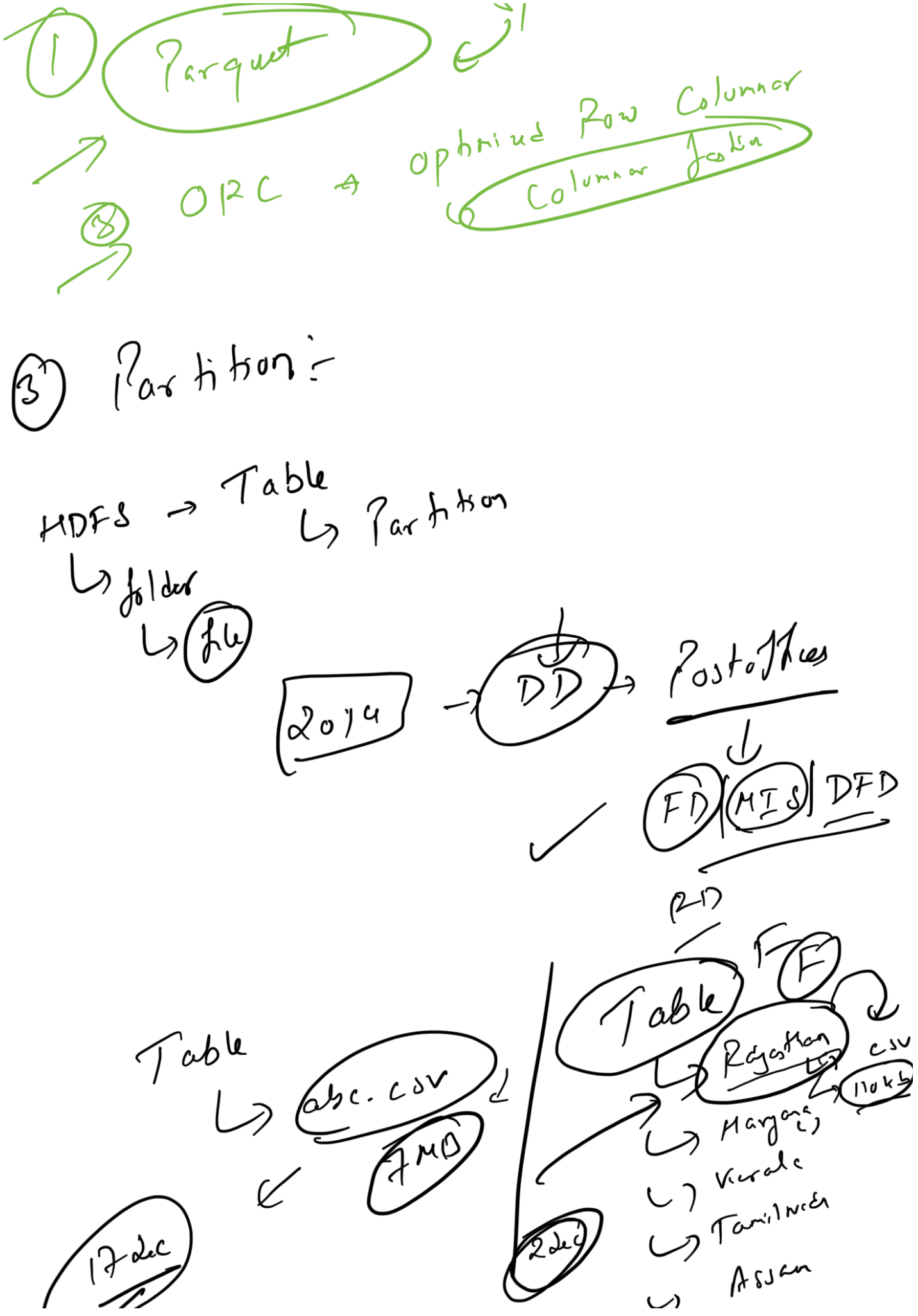
(2) Parquet file
↳ Columnar storage



max(sal) → 600

Select * from emp;

Spark


1
Parquet
2
ORC → optimized Row Columnar
Columnar format
3
Partition :-
HDFS → Table
↳ Partition
↳ folder
↳ file
2014
DD
Postoffices
FD
MIS
DFD
Table
abc.csv
FMD
17dec
2dec
Table
F
Rajasthan
csv
110kb
↳ Haryana
↳ Kerala
↳ TamilNadu
↳ Assam