Big Data Technologies
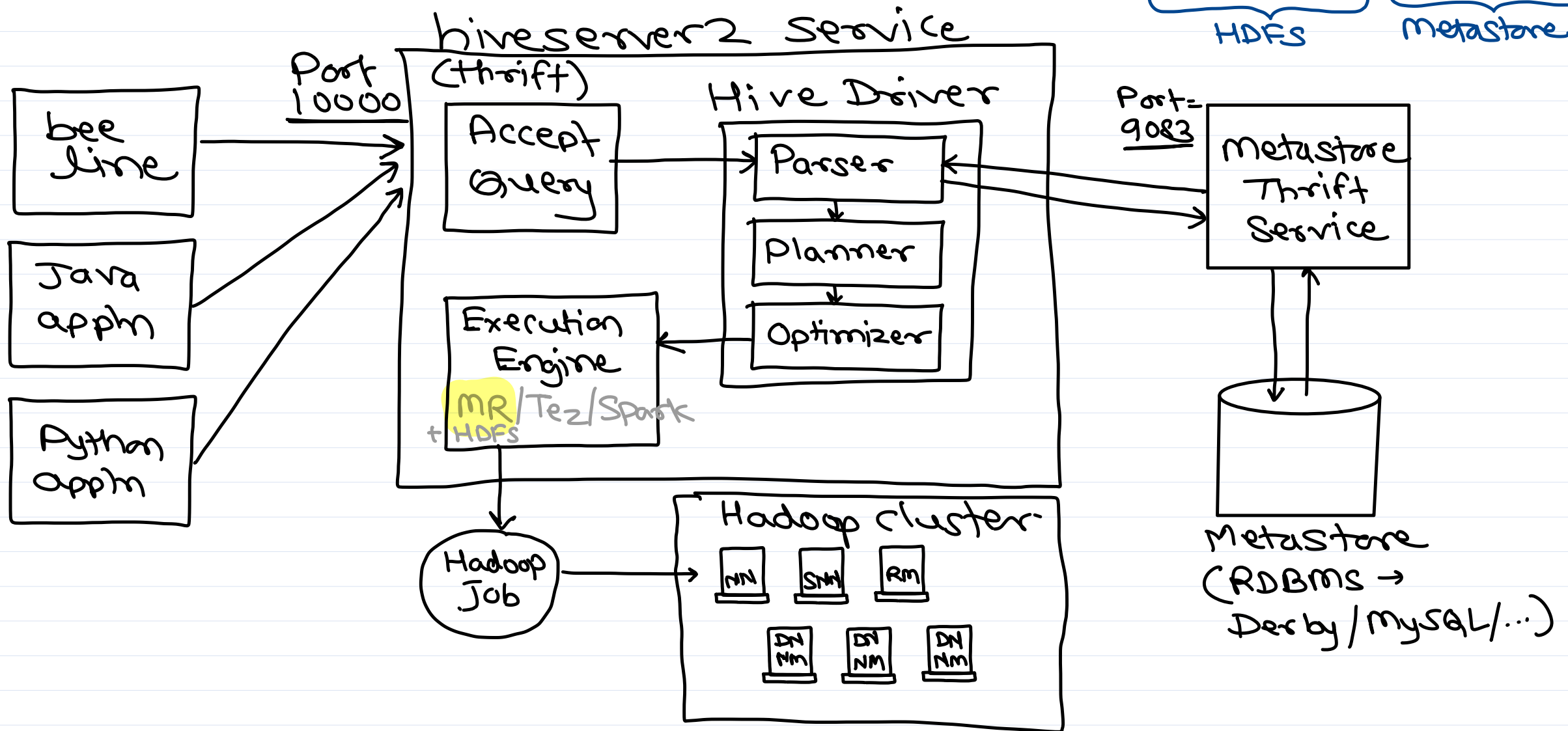
Trainer: Mr. Nilesh Ghule.

Hive Architecture

Hive is a RDBMS → Table ; Table = Data + Structure
                    → Views ...   (Rows|Cols)   (Cols)
                                          HDFS   Metastore

hiveserver2 service (thrift)

Port 10000

bee line → (hiveserver2 service)

Java appln →

Python appln →

Accept Query → **Parser**

Hive Driver
- Parser
- Planner
- Optimizer

Execution Engine
MR/Tez/Spark
+ HDFS

→ Optimizer → Execution Engine

Hadoop Job →

Hadoop cluster
NN SNN RM

DN/NM DN/NM DN/NM

Port = 9083

Metastore Thrift Service

Metastore (RDBMS → Derby / MySQL / ...)

Data Formats and Communication Protocols

Data Formats → Storage.   Communication → Data transfer

Text Formats   Text Formats
- CSV   + more human readable   - JSON ← REST api
- JSON   - unefficient for   - XML ← web services
- XML   parsing.
...   ...

Binary Formats   Binary Formats
- ORC   + very efficient   - Thrift (Apache)
- Parquet   - not human   - Protobuf (Google)
- Avro   readable.   - AVRO (Apache)
- Sequence File + Designed to work
- BSON   irrespective of tech/platform.
...   ↳ Data type support
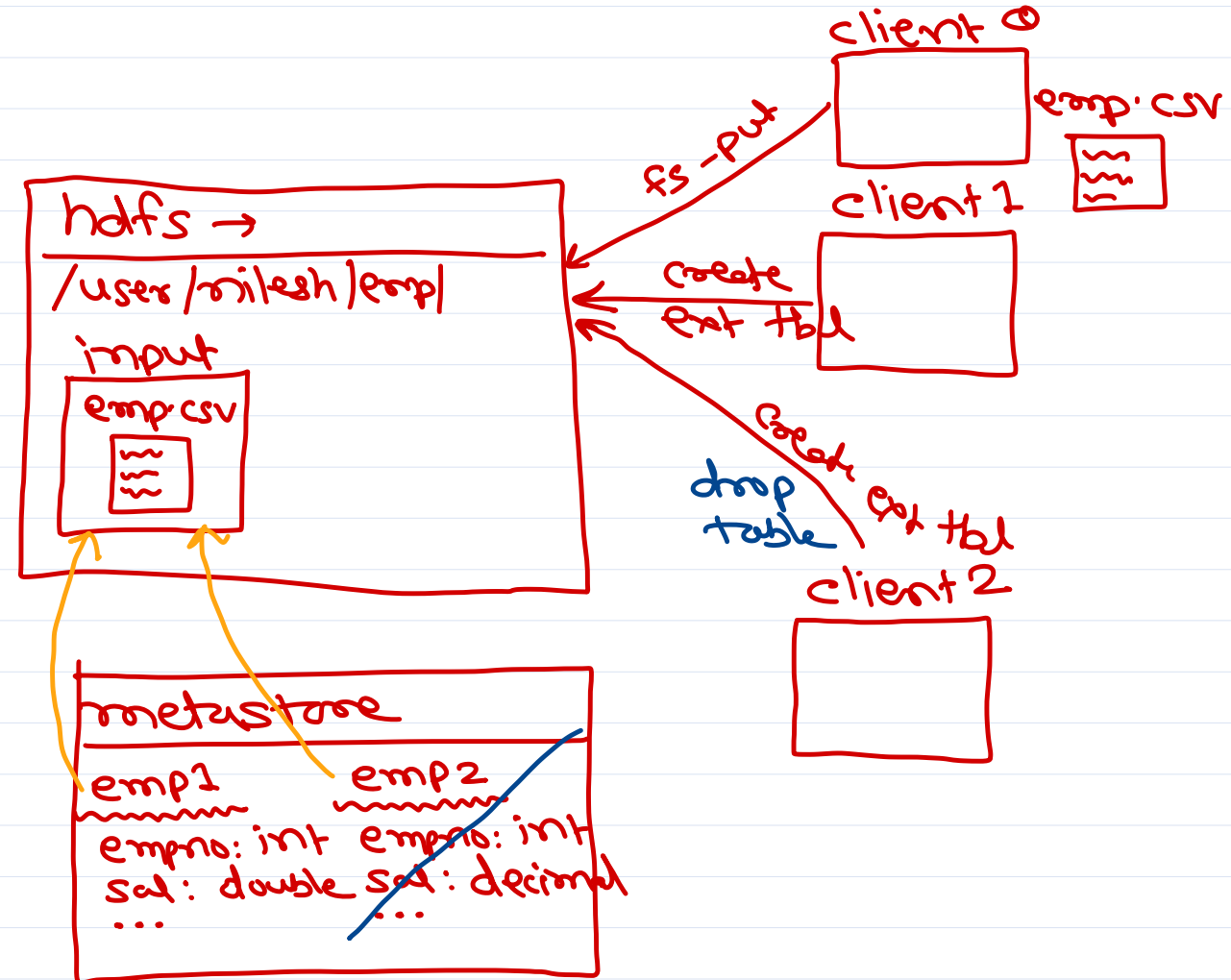
Managed Table vs External Table

managed tables: emp1, emp2
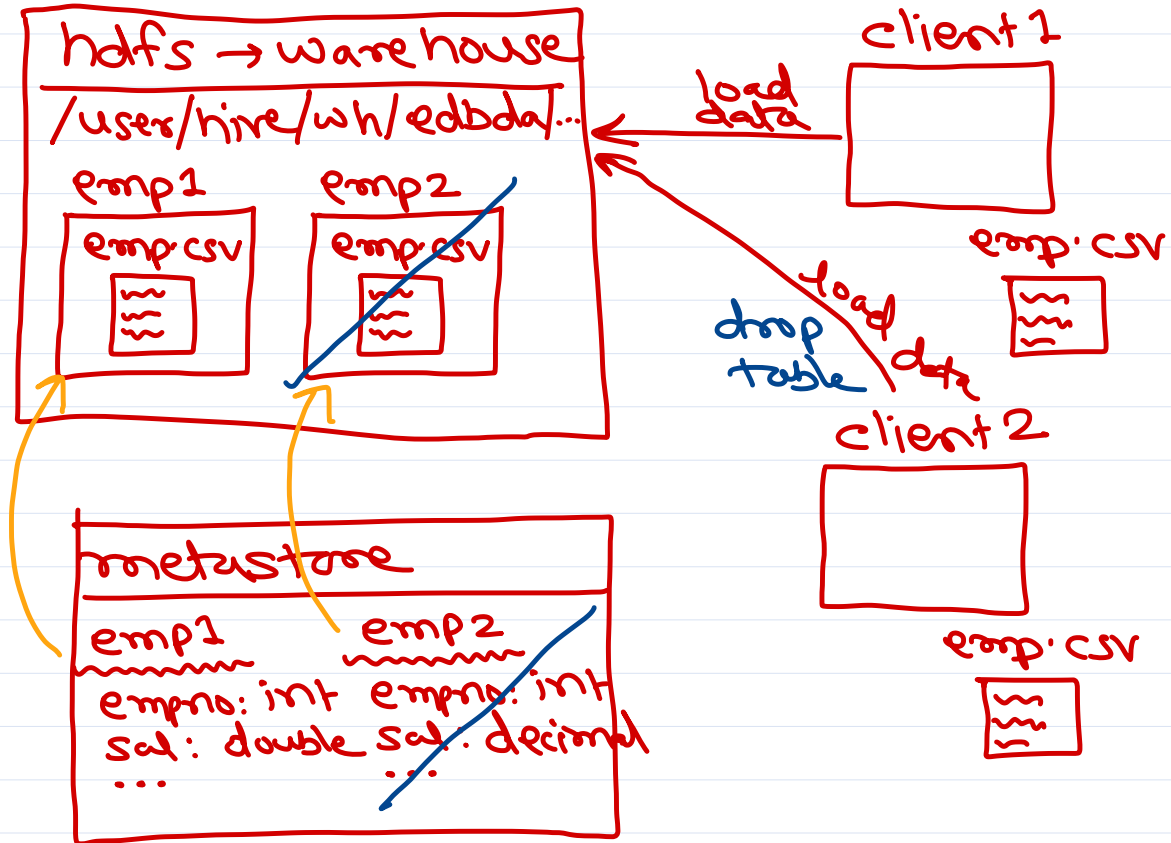
external tables: emp1, emp2

hdfs → warehouse

/user/hive/wh/edbda/...

emp1

emp.csv

emp2

emp.csv

load data

client 1

emp.csv

drop table

load data

client 2

emp.csv

metastore

emp1   emp2

empno: int empno: int

sal: double sal: decimal

...   ...

hdfs →

/user/nilesh/emp1

input

emp.csv

client 0

emp.csv

fs -put

client 1

emp.csv

create ext tbl

create ext tbl

drop table

client 2

metastore

emp1   emp2

empno: int empno: int

sal: double sal: decimal

...   ...

Thank you!

Nilesh Ghule <nilesh@sunbeaminfo.com>