APACHE HIVE

Part 1

by Sumit Mittal

IMPORTANT

Copyright Infringement and Illegal Content Sharing Notice

Trainer Introduction

Mr. Sumit Mittal, CEO & founder of **TrendyTech**. He has a Master's degree in Computer Applications from NIT Trichy & have a total of 7+ years of industry experience. He has worked for top MNC's like **Cisco** & **VMware**.

Consistent 5 star Google rated Bigdata course

Transactional and Analytical Processing

Transactional and Analytical Processing

Order Management Support

Michael is responsible for tracking and delivering orders on time

Revenue Analyst

Emma is responsible for tracking and monitoring revenues

Order Management Support

20 deliveries in Kent, WA are delayed

Michael is responsible for tracking and delivering orders on time

The courier company has had a computer outage

Order Management Support

3 customers want to ship orders to a different address

Michael updates the address on the shipments and re-routes them

Revenue Analyst

Her manager wants an update on last month's revenues

Last month was an unusually slow one

Emma pulls up data for the last 5 years to check for seasonal effects

Revenue Analyst

Management asks if the new TV ads this year were successful

Emma checks customer signups for a jump when the campaigns were run

Transactional and Analytical Processing

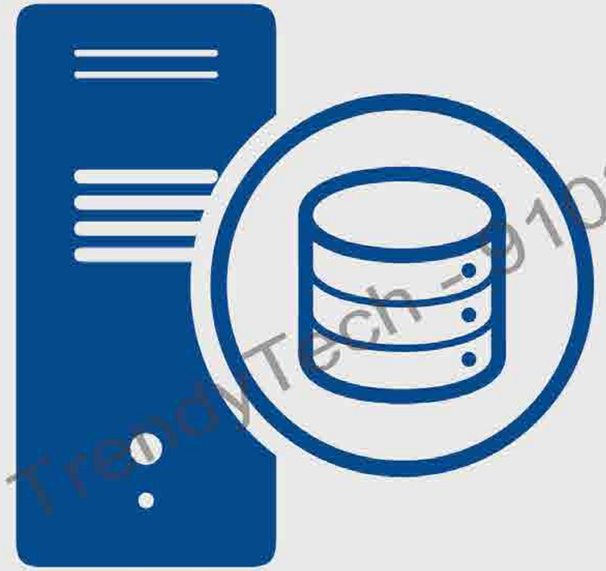
Transactional Processing

Analytical Processing

Transactional and Analytical Processing

Transactional Processing	Analytical Processing
Analyzes individual entries	Analyzes large batches of data
Access to recent data, from thelast few hours or days	Access to older data going back months, or even years
Updates data	Only reads data
Fast real-time access	Long running jobs
Usually a single data source	Multiple data sources

Transactional and Analytical Processing



Small Data

Both these objectives could be achieved using the same database system

Small Data

Single machine with backup

Structured, well-defined data

Can access individual records or the entire dataset

No replication, updated data available instantaneously

Different tables store data from different sources

Transactional and Analytical Processing

BIG Data

Very hard to meet all requirements with
the same database system

Big Data

Data distributed on a cluster with multiple machines

Semi-structured or unstructured data

No random access to data

Data replicated, propagation of updates take time

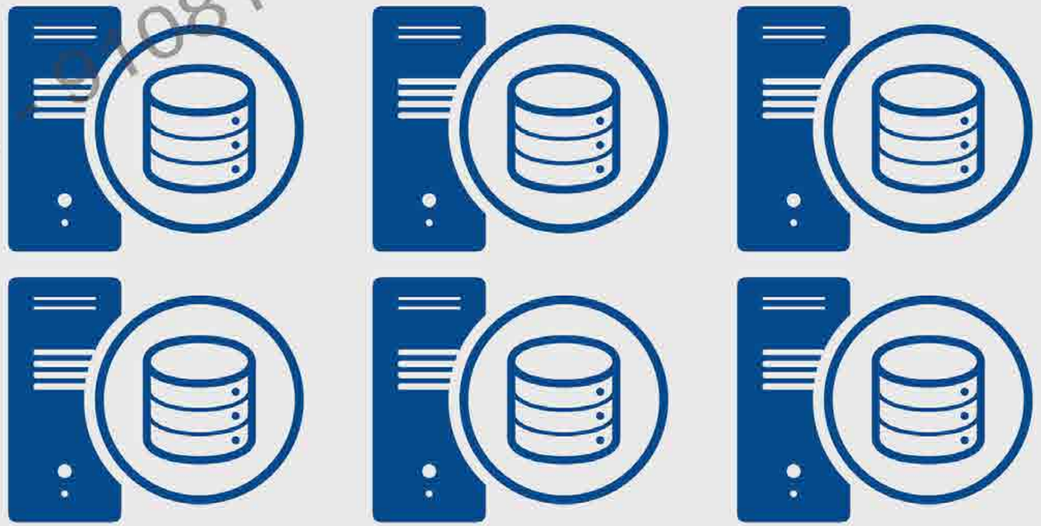
Different sources may have different unknown formats

The same infrastructure cannot support both **transactional** and **analytical** processing

Transactional and Analytical Processing

Transactional Processing

Analytical Processing



Traditional RDBMS

Data Warehouse

Data Warehouse for Analytical Processing

Data Warehouse

Long running batch jobs

Optimized for read operations

Holds data from multiple sources

Holds data over a long period of time

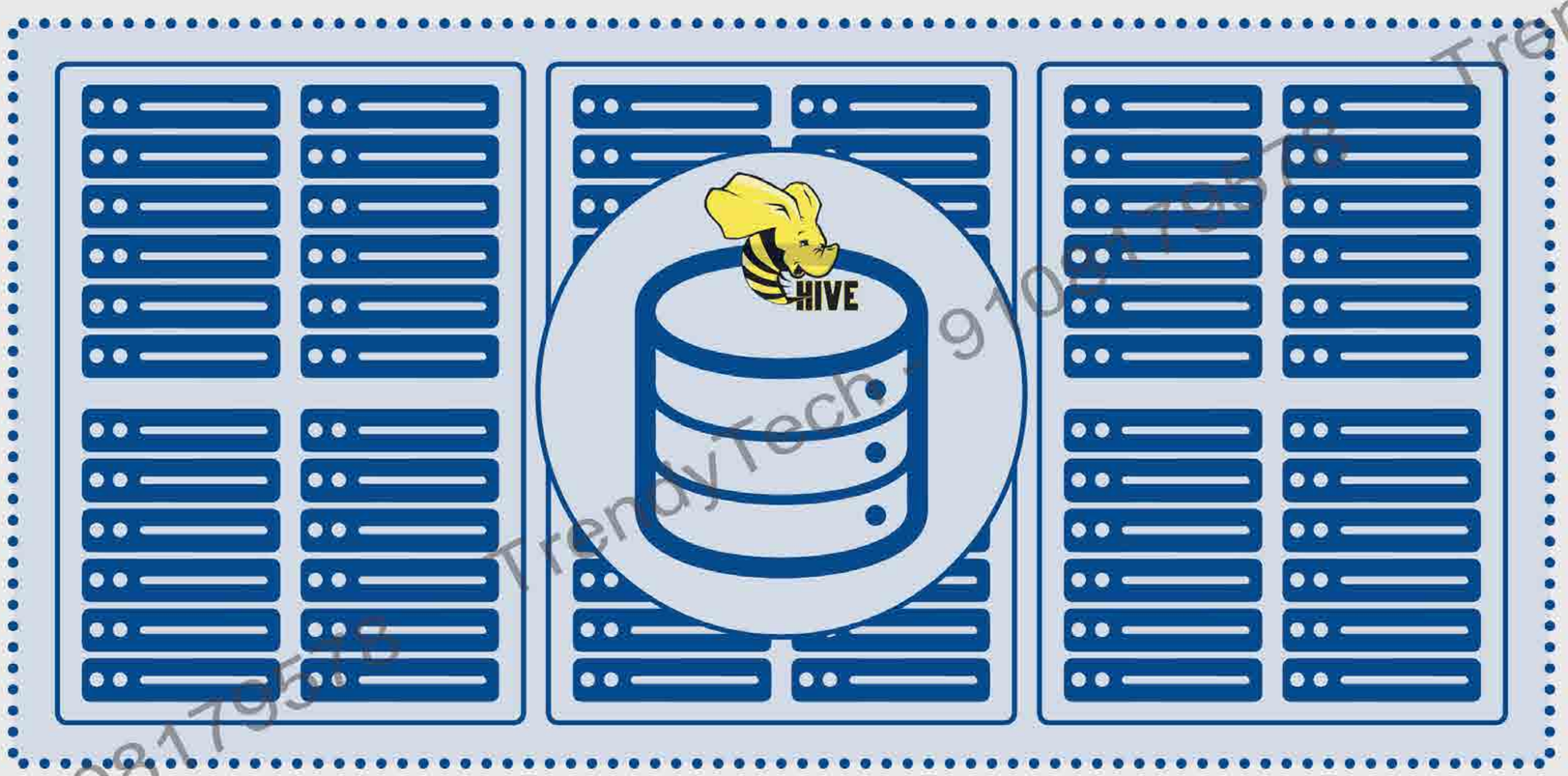
Data may be lagged, not real-time

Data Warehouse

Examples of data warehouses

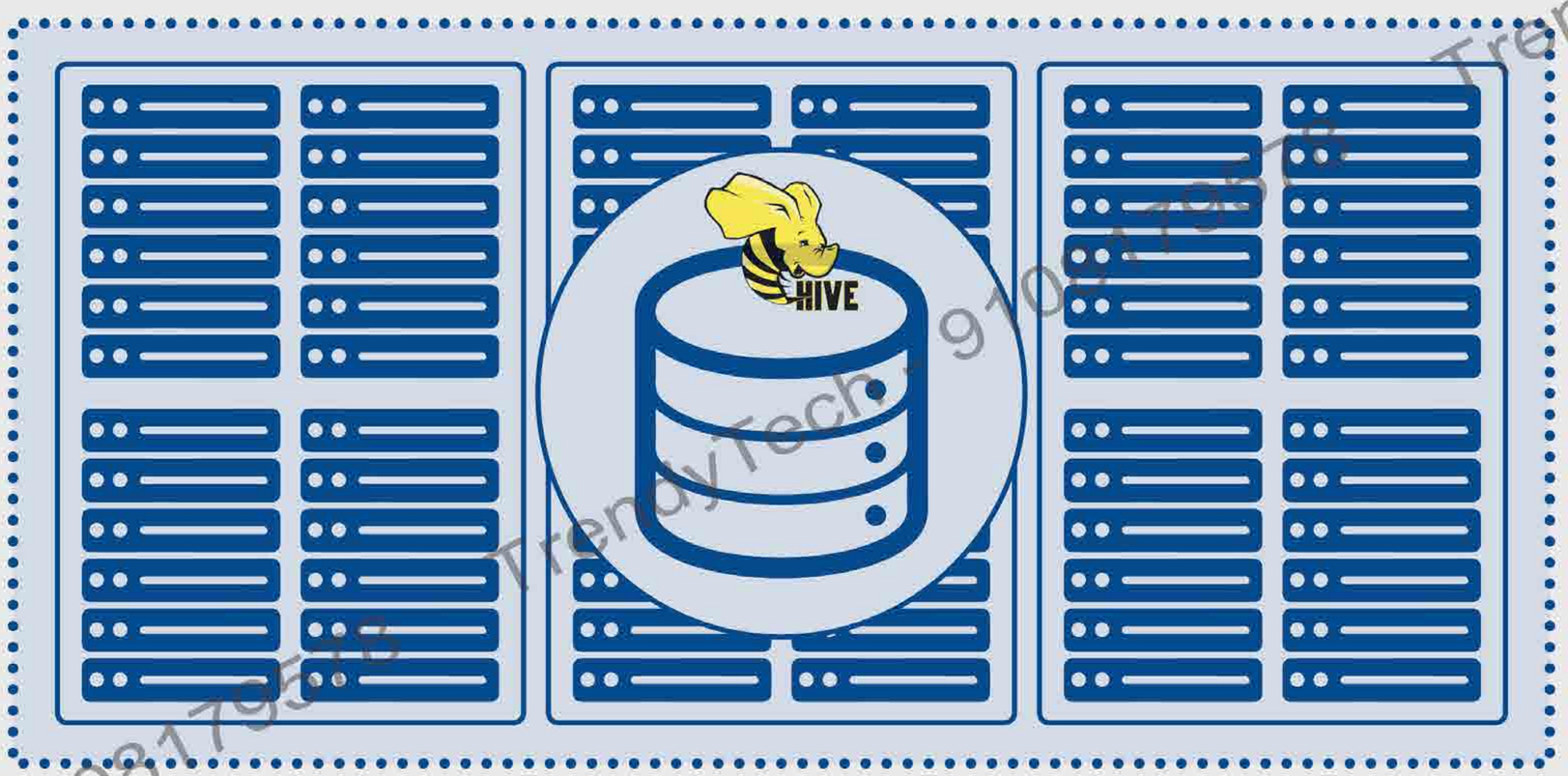
- Vertica
- Teradata
- Oracle
- IBM

Data Warehouse



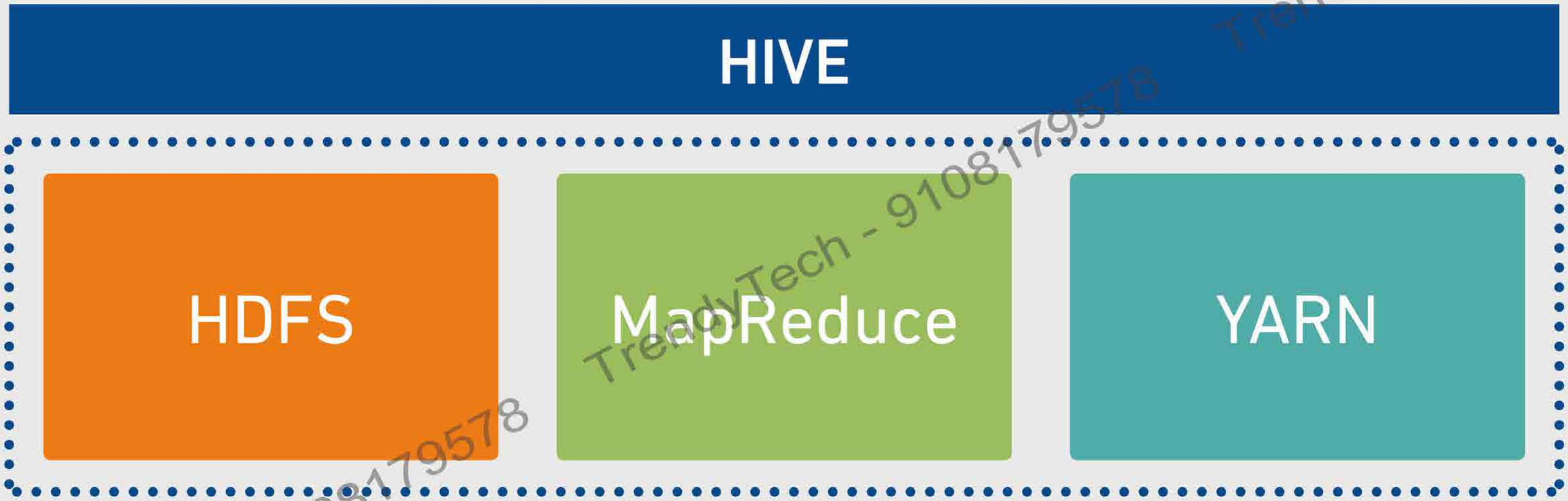
Apache Hive is an open-source data warehouse

Data Warehouse



Hive is part of the larger Hadoop ecosystem

Hive on Hadoop

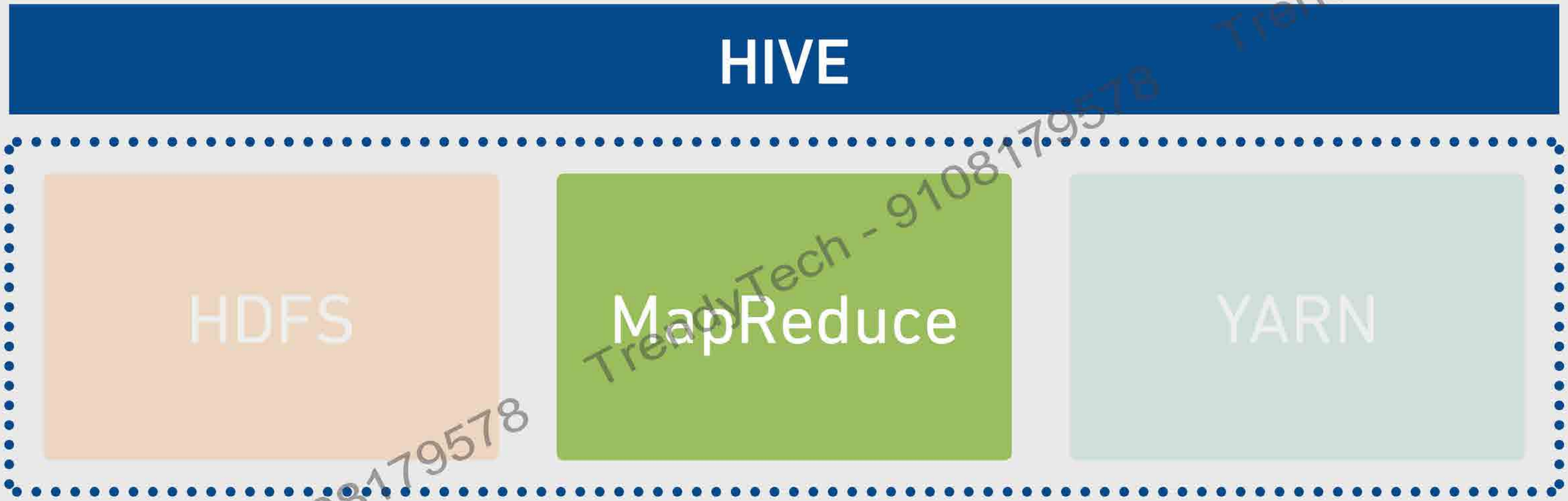


Hive runs on top of the Hadoop distributed computing framework

Hive on Hadoop

Hive stores its data in HDFS

Hive on Hadoop



Hive runs all processes in the form of MapReduce jobs under the hood

MapReduce

MapReduce

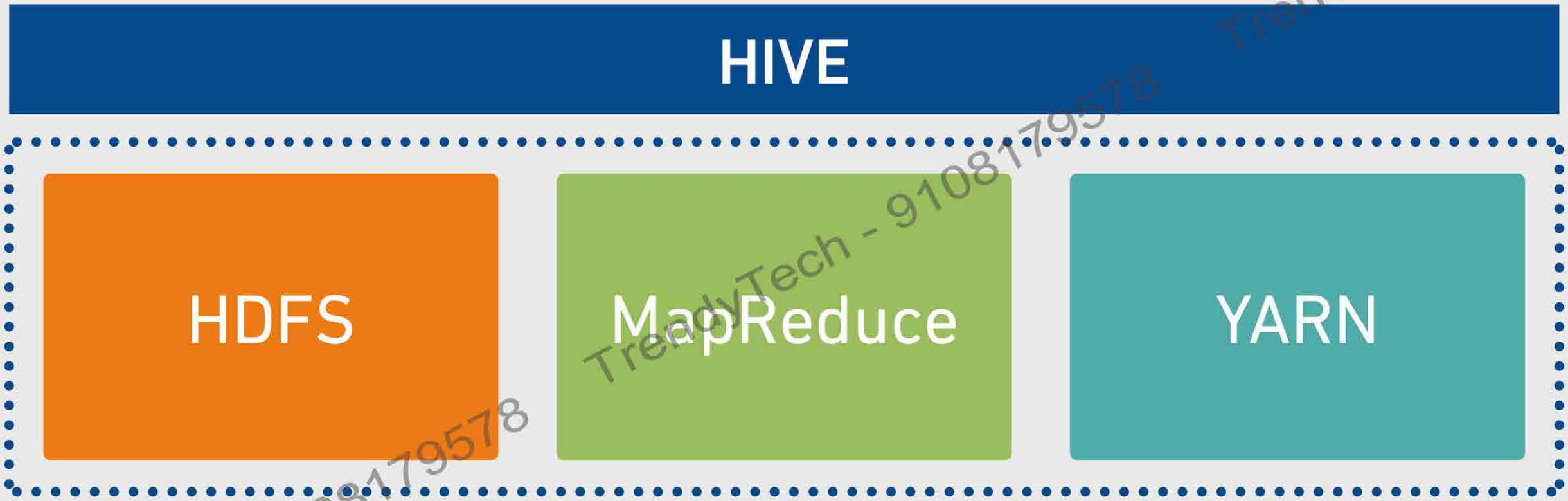
A parallel programming model

Defines the logic to process data on multiple machines

Batch processing operations on files in HDFS

Usually written in Java using the Hadoop MapReduce library

Hive on Hadoop



Do we have to write MapReduce code to work with Hive ?

No

HiveQL

Hive Query Language

A SQL-like interface to the underlying data

HiveQL

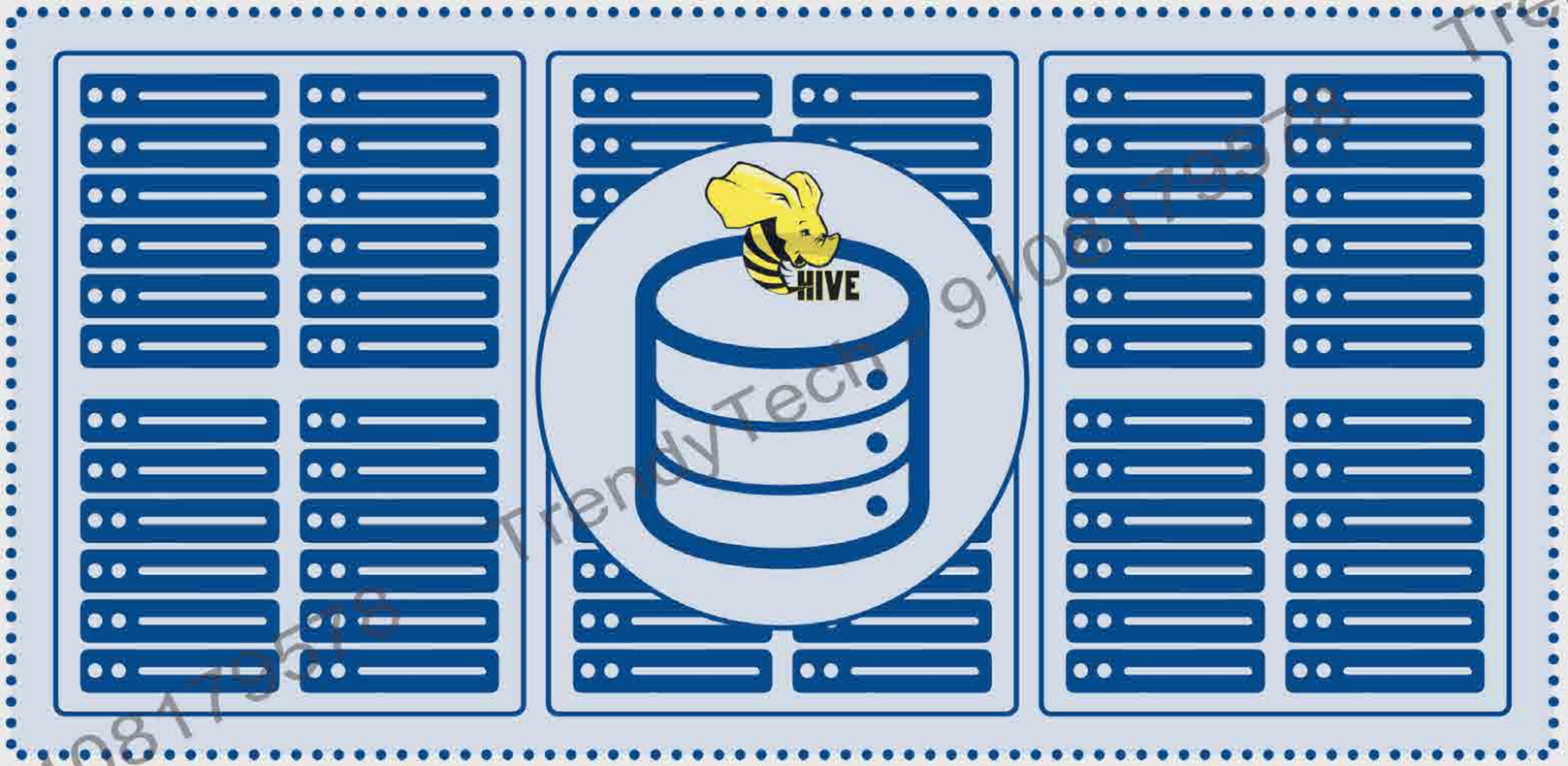
Modeled on the Structured Query Language (SQL)

Familiar to analysts and engineers

Simple query constructs

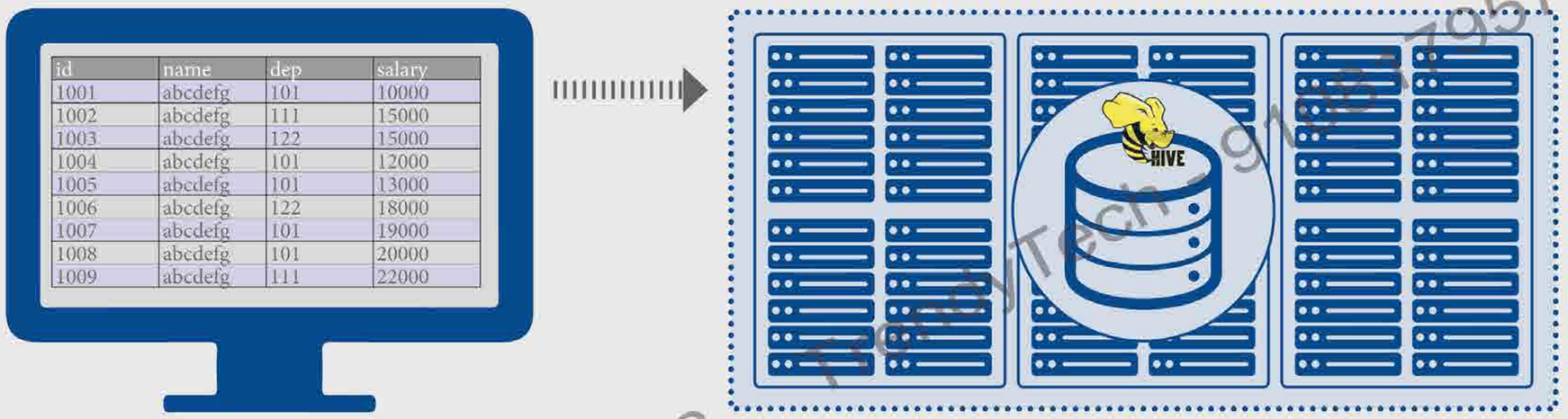
- select
- group by
- join

HiveQL



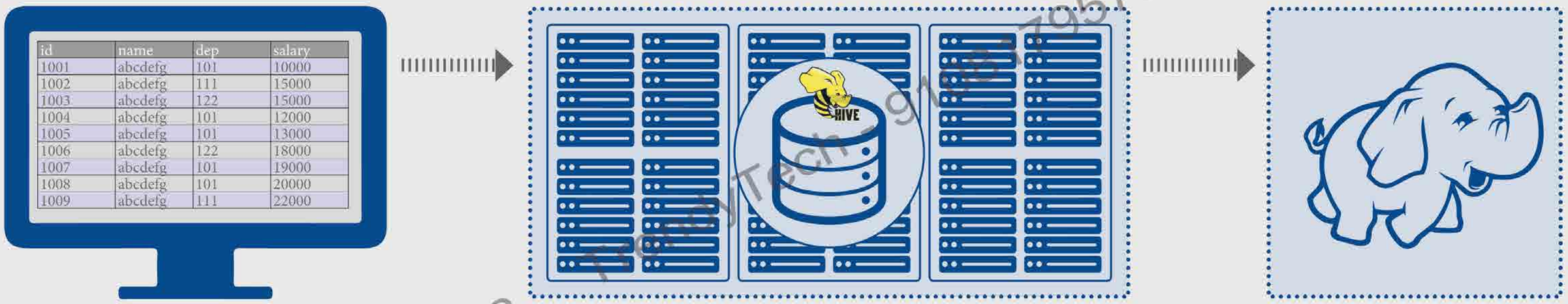
Hive exposes files in HDFS in the form of tables to the user

HiveQL



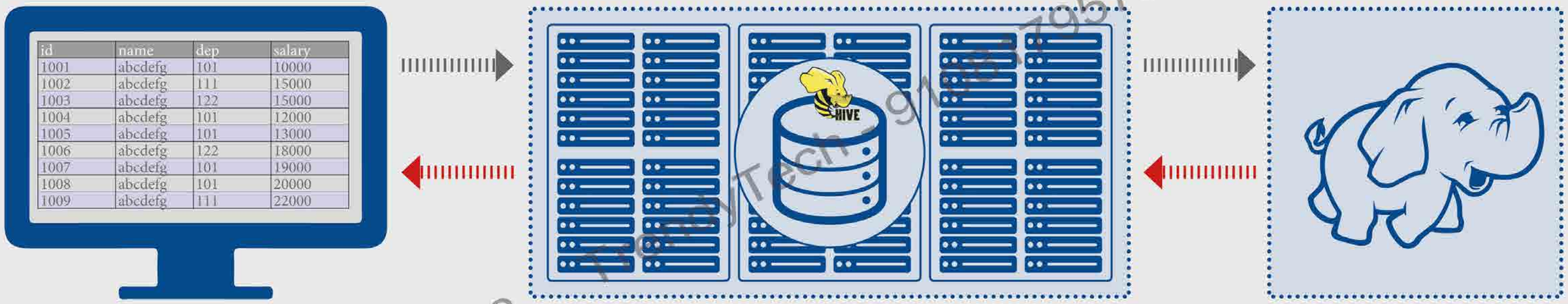
Hive exposes files in HDFS in the form of tables to the user

HiveQL



Hive will translate the query to MapReduce tasks and run them on Hadoop

HiveQL



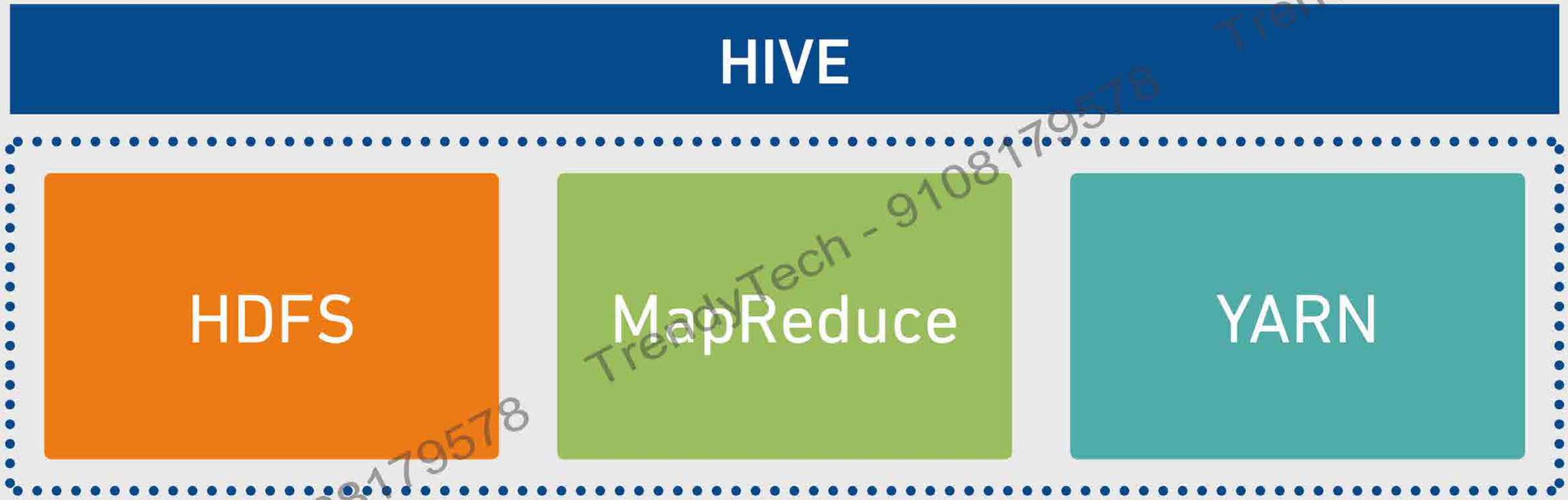
MapReduce will process files on HDFS and return results to Hive

Hive **abstracts away** the details of the underlying MapReduce jobs

Work with Hive **almost** exactly like you would with a traditional database

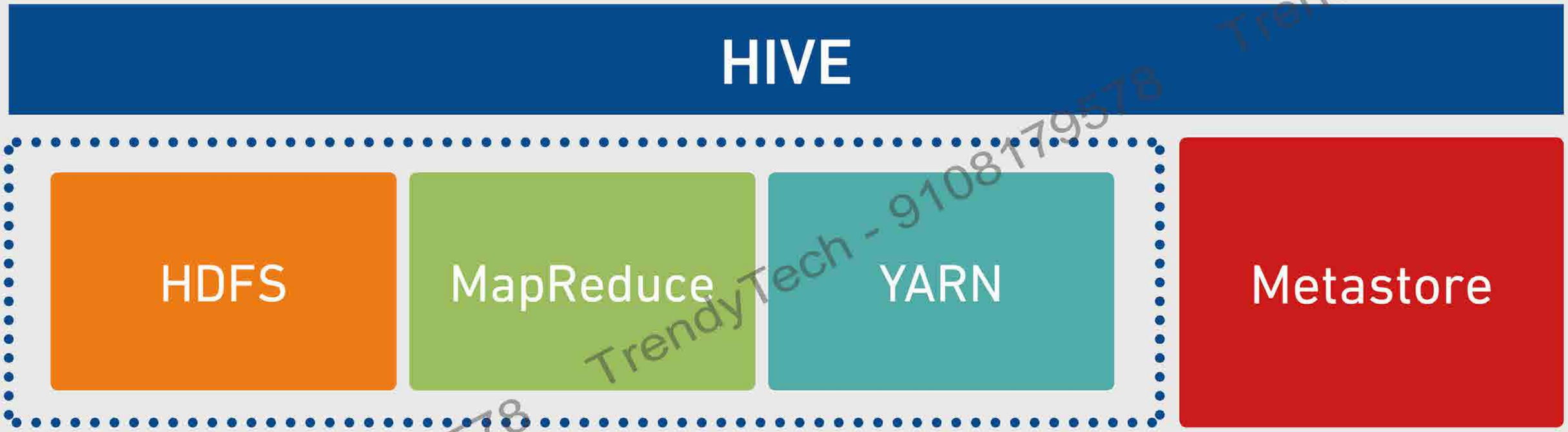
The Hive Metastore

The Hive Metastore



A Hive user sees data as if they were stored in tables

The Hive Metastore



Exposes the file-based storage of HDFS in the form of tables

The Hive Metastore

Metastore

The bridge between data stored in files and the tables exposed to users

The Hive Metastore

Stores metadata for all the tables in Hive

Maps the files and directories in Hive to tables

Holds table definitions and the schema for each table

Has information on converting files to table representations

The Hive Metastore

Any database with a JDBC driver can be used as a metastore

The Hive Metastore

Development environments use the built-in Derby database

Embedded metastore

Hive vs. RDBMS

Hive vs. RDBMS

Hive vs. RDBMS

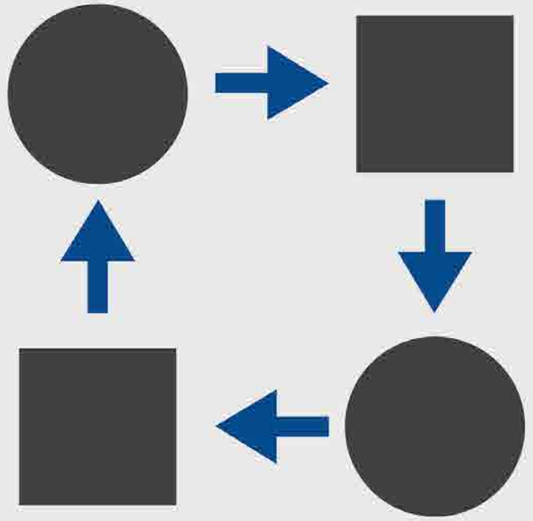
Hive	RDBMS
Large datasets	Small datasets
Parallel computations	Serial computations
High latency	Low latency
Read operations	Read/write operations
Not ACID compliant by default	ACID compliant
HiveQL	SQL

Hive vs. RDBMS

Large datasets

Calculating trends

Small datasets



Accessing and updating individual records

Hive vs. RDBMS

Parallel Computations

Distributed system with multiple machines

Serial Computations

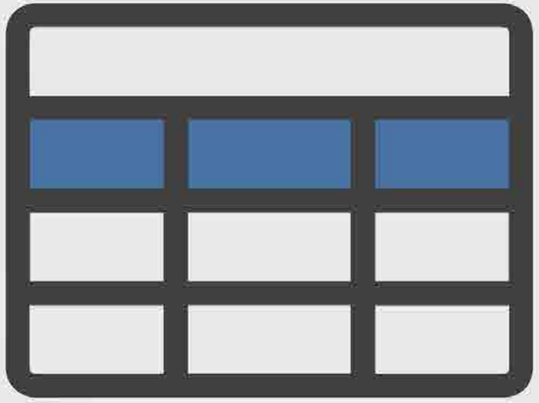
Single computer with backup

Hive vs. RDBMS

Parallel Computations

Semi-structured data files partitioned across machines

Serial Computations



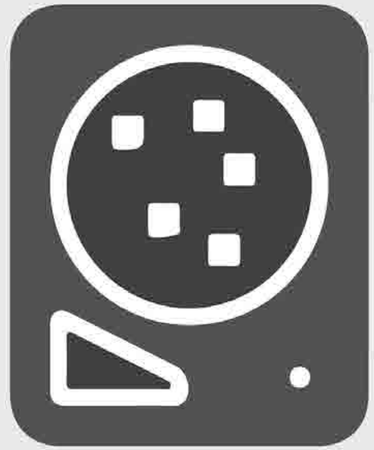
Structured data in tables on one machine

Hive vs. RDBMS

High Latency

Semi-structured data files partitioned across machines

Low Latency



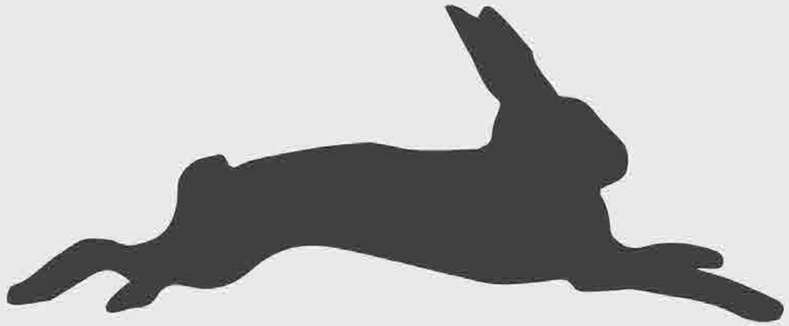
Structured data in tables on one machine

Hive vs. RDBMS

High Latency

Fetching a row will run a MapReduce that might take minutes

Low Latency



Queries can be answered in milliseconds or microseconds

Hive vs. RDBMS

Read Operations

Schema-on-read

Read/Write Operations

Schema-on-write

Hive vs. RDBMS

Read Operations

Data can be dumped into Hive tables from any source

Read/Write Operations

Only data which satisfies constraints are stored in the database

Hive vs. RDBMS

HQL	SQL
Schema on read, no constraints enforced	Schema on write keys, not null, unique all enforced
Minimal index support	Indexes allowed
Row level updates, deletes as a special case	Row level operations allowed in general
Many more built-in functions	Basic built-in functions
Only equi-joins allowed	No restriction on joins
Restricted subqueries	Whole range of subqueries

Google REVIEWS

5 Star Google Rated Big Data Course

LEARN FROM THE EXPERT

9 1 0 8 1 7 9 5 7 8

Call for more details

Follow US

Trainer	Mr. Sumit Mittal
Phone	9108179578
Email	trendytech.sumit@gmail.com
Website	https://trendytech.in/courses/big-data-online-training/
Linked In	https://www.linkedin.com/in/bigdatabysumit/
Twitter	@BigdataBySumit
Instagram	bigdatabysumit
Facebook	https://www.facebook.com/trendytech.in/
Youtube	TrendyTech