INTRO TO HIVE

FRED AND GEORGE BOTH WORK AT AN E-COMMERCE COMPANY

FRED WORKS FOR THE REVENUE ANALYTICS TEAM

GEORGE WORKS FOR THE ORDER MANAGEMENT TEAM

FRED
REVENUE ANALYTICS

GEORGE
ORDER MANAGEMENT

HERE'S WHAT THEIR DAILY ROUTINES LOOK LIKE

FRED

REVENUE ANALYTICS

9:00 AM

SENDS AN EMAIL TO EACH CATEGORY MANAGER WITH CATEGORY LEVEL REVENUE

FRED

REVENUE ANALYTICS

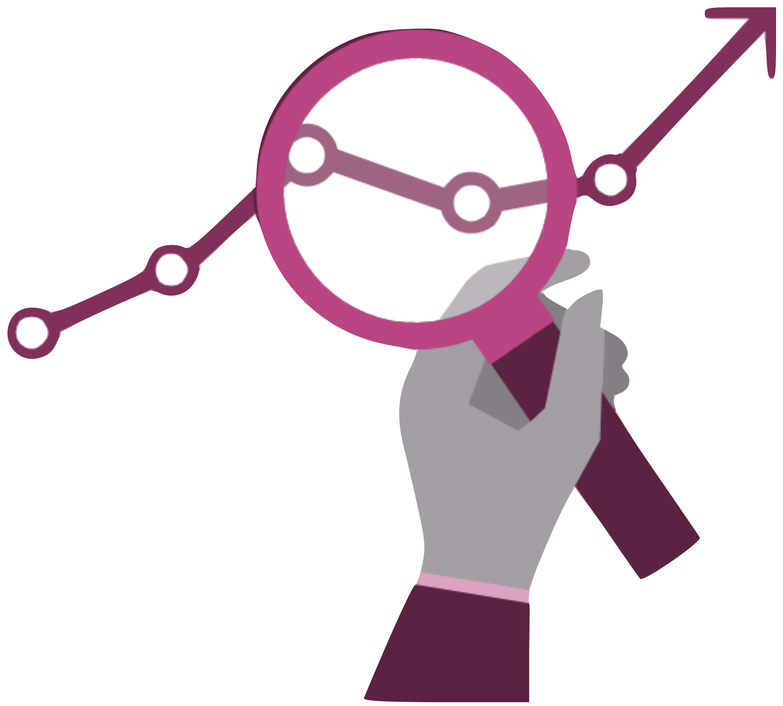
10:00 AM

STARTS DOING AN ANALYSIS
OF LAST MONTH OF SALES

FRED

REVENUE ANALYTICS

11:00 AM



FINDS AN ANOMALOUS DROP IN SALES FROM LAST MONTH

FRED

REVENUE ANALYTICS

11:00 AM - 3:00 PM

PULLS THE REVENUE DATA FOR LAST 3 YEARS TO SEE IF THE ANOMALY OCCURRED IN THE PAST

FRED

REVENUE ANALYTICS

11:00 AM - 3:00 PM

THIS TAKES A WHILE AS THE ANALYSIS JOB NEEDS TO RUN ON LOT'S OF DATA

FRED

REVENUE ANALYTICS

3:00 PM -
5:00 PM

PREPARES A PRESENTATION WITH
LAST WEEK'S REVENUE PERFORMANCE

FRED
REVENUE ANALYTICS

5:00 PM-7:00 PM

PRESENTS THE DECK TO MANAGEMENT AT A WEEKLY REVENUE MEETING

FRED

REVENUE ANALYTICS

5:00 PM-7:00 PM

THERE ARE A BUNCH OF QUESTIONS

WHAT KIND OF EFFECT DID LAST WEEK'S PROMOTION HAVE ?

THERE'S BEEN A BUNCH OF COMPLAINTS ON TWITTER. HAS THERE BEEN ANY EFFECT ON SALES?

FRED

REVENUE ANALYTICS

7:00 PM-9:00 PM

SETS UP NEW ANALYSIS JOBS TO RUN OVERNIGHT BASED ON THE QUESTIONS ASKED AT THE MEETING

FRED

REVENUE ANALYTICS

9:00 PM

SPOKE TO THE BOSS ABOUT THE ANOMALY HE FOUND DURING THE DAY

BOSS WANTS A DECK WITH ALL THE FINDINGS BY MORNING

FRED

REVENUE ANALYTICS

12:00 AM

FINISHED AND MAILED OFF THE DECK TO THE BOSS

CHECKS ON THE JOBS RUNNING OVERNIGHT

FRED

REVENUE ANALYTICS

12:30 AM

TRIGGERS THE JOB FOR DAILY CATEGORY EMAIL REPORT TO BE SENT BY MORNING

GOES TO SLEEP!

GEORGE
ORDER MANAGEMENT

9:00 AM

GEORGE OPENS UP THE ORDER MANAGEMENT PORTAL

GEORGE
ORDER MANAGEMENT

9:00 AM

THE PORTAL HAS REAL-TIME STATUS OF ALL ORDERS IN THE SYSTEM

GEORGE
ORDER MANAGEMENT

9:00 AM

SOMETIMES PENDING DELIVERIES
WITH A RISK OF DELAY ARE
ASSIGNED TO AN AGENT LIKE GEORGE

GEORGE
ORDER MANAGEMENT

9:00 AM

GEORGE GETS AN UPDATE ON WHAT ORDERS HE IS RESPONSIBLE FOR GETTING DELIVERED TODAY

GEORGE
ORDER MANAGEMENT

10:00 AM

HE HAS CALLED UP EACH DELIVERY PERSON TO FIND OUT THE STATUS OF EACH ORDER

GEORGE
ORDER MANAGEMENT

11:00 AM

BASED ON STATUS UPDATES, GEORGE ASSESSED THE RISK OF DELAY OF EACH OF THE ORDERS

GEORGE
ORDER MANAGEMENT

11:00 AM - 5:00 PM

HE CALLS UP CUSTOMERS WHOSE ORDER WILL DEFINITELY BE DELAYED

GEORGE
ORDER MANAGEMENT

11:00 AM -
5:00 PM

SOME CUSTOMERS WANT TO
CANCEL

GEORGE

ORDER MANAGEMENT

11:00 AM - 5:00 PM

CANCEL

SOME ARE OK WITH A DELAYED DELIVERY IF THEY GET AN ADDITIONAL DISCOUNT

GEORGE

ORDER MANAGEMENT

11:00 AM - 5:00 PM

CANCEL

ADDITIONAL DISCOUNT

SOME WANT TO ORDER RE-ROUTED TO A DIFFERENT ADDRESS

GEORGE
ORDER MANAGEMENT

11:00 AM -
5:00 PM

GEORGE MAKES
ALL THESE UPDATES
TO THE SYSTEM

CANCEL

ADDITIONAL DISCOUNT

RE-ROUTE

GEORGE
ORDER MANAGEMENT

11:00 AM -
5:00 PM

EACH UPDATE HAS
AN IMMEDIATE
CASCADING EFFECT
IN THE SUPPLY CHAIN

CANCEL

ADDITIONAL DISCOUNT

RE-ROUTE

GEORGE
ORDER MANAGEMENT

11:00 AM - 5:00 PM

ALL THESE NEED TO HAPPEN NEARLY INSTANTANEOUSLY

CANCEL

REVERSES THE ORDER PATH FROM CUSTOMER TO WAREHOUSE

ADDITIONAL DISCOUNT

UPDATES THE CUSTOMER'S CREDIT IN THE ACCOUNTING SYSTEM

RE-ROUTE

A NEW DELIVERY ROUTE AND DELIVERY PERSON ARE ASSIGNED

GEORGE

ORDER MANAGEMENT

5:00 PM-7:00 PM

GEORGE GETS NOTIFIED OF AN URGENT OVERNIGHT DELIVERY

THE REGULAR VENDOR SEEMS TO HAVE RUN OUT OF STOCK

GEORGE
ORDER MANAGEMENT

5:00 PM-7:00 PM

GEORGE IDENTIFIES ANOTHER VENDOR
WHO HAS THE STOCK AVAILABLE

UPDATES THE ORDER DELIVERY PLAN
TO USE THE NEW VENDOR

GEORGE
ORDER MANAGEMENT

7:00 PM-9:00 PM

FOLLOWS UP ON ALL THE ORDERS THAT WERE MEANT TO BE DELIVERED TODAY

MARKS ANY UNDELIVERED ORDERS FOR THE NEXT SHIFT TO HANDLE

GEORGE
ORDER MANAGEMENT

9:00 PM

LOGS OUT AND HEADS HOME

FRED

REVENUE ANALYTICS

GEORGE

ORDER MANAGEMENT

LET'S QUICKLY SUMMARIZE

FRED

REVENUE ANALYTICS

ANALYZES BATCHES OF ORDERS

CATEGORY WISE, MONTHLY, WEEKLY, YEARLY

GEORGE

ORDER MANAGEMENT

ANALYZES INDIVIDUAL ORDERS

FRED

REVENUE ANALYTICS

BATCHES

ONLY READS THE ORDER DATA

GEORGE

ORDER MANAGEMENT

INDIVIDUAL ORDERS

KEEPS UPDATING THE ORDER DATA

FRED

REVENUE ANALYTICS

BATCHES

READ

IS OK WITH LONG -RUNNING JOBS

GEORGE

ORDER MANAGEMENT

INDIVIDUAL ORDERS

READ /UPDATE

NEEDS EVERYTHING REAL-TIME / INSTANTANEOUSLY

FRED

REVENUE ANALYTICS

BATCHES

READ

LONG-RUNNING JOBS

NEEDS TO BE ABLE TO CONNECT THE DOTS BETWEEN MULTIPLE TYPES OF DATA

PROMOTIONS, CUSTOMER COMMUNICATIONS, ORDERS ETC

GEORGE

ORDER MANAGEMENT

INDIVIDUAL ORDERS

READ /UPDATE

INSTANTANEOUS

JUST NEEDS TO SEE THE CURRENT ORDERS DATA

FRED
REVENUE ANALYTICS

BATCHES

READ

LONG-RUNNING JOBS

MULTIPLE DATA SOURCES

GEORGE
ORDER MANAGEMENT

INDIVIDUAL ORDERS

READ /UPDATE

INSTANTANEOUS

SINGLE DATA SOURCE

BOTH FRED AND GEORGE FREQUENTLY NEED TO LOOK AT ORDERS DATA

FRED

REVENUE ANALYTICS

BATCHES

READ

LONG-RUNNING JOBS

MULTIPLE DATA SOURCES

GEORGE

ORDER MANAGEMENT

INDIVIDUAL ORDERS

READ /UPDATE

INSTANTANEOUS

SINGLE DATA SOURCE

BOTH OF THESE GUYS NEED A DATABASE THAT STORES THE DATA IN THE FORM OF TABLES

FRED

REVENUE ANALYTICS

BATCHES

READ

LONG-RUNNING JOBS

MULTIPLE DATA SOURCES

GEORGE

ORDER MANAGEMENT

INDIVIDUAL ORDERS

READ /UPDATE

INSTANTANEOUS

SINGLE DATA SOURCE

BUT THE DATABASE WILL NEED TO SERVE VERY DIFFERENT NEEDS

FRED

REVENUE ANALYTICS

BATCHES

GEORGE

ORDER MANAGEMENT

INDIVIDUAL ORDERS

READ

READ /UPDATE

LONG-RUNNING

INSTANTANEOUS

MULTIPLE

SINGLE DATA SOURCE

SOURCES

FRED'S DATABASE NEEDS TO PROCESS LARGE AMOUNTS OF DATA AT A TIME

FRED

REVENUE ANALYTICS

BATCHES

GEORGE

ORDER MANAGEMENT

INDIVIDUAL ORDERS

READ

READ /UPDATE

LONG-RUNNING

MISCELLANEOUS

MULTIPLE DATA SOURCES

SINGLE DATA SOURCE

GEORGE'S DATABASE NEEDS TO ACCESS ONLY 1 ORDER AT A TIME

FRED

REVENUE ANALYTICS

BATCHES

READ

LONG-RUNNING JOBS

MULTIPLE DATA SOURCES

GEORGE

ORDER MANAGEMENT

INDIVIDUAL ORDERS

READ /UPDATE

INSTANTANEOUS

SINGLE DATA SOURCE

FRED'S DATABASE WILL NEED TO JUST WRITE THE DATA ONCE, THEN ONWARDS WE ONLY READ THE DATA

FRED

REVENUE ANALYTICS

BATCHES

READ

LONG-RUNNING JOBS

GEORGE

ORDER MANAGEMENT

INDIVIDUAL ORDERS

READ /UPDATE

INSTANTANEOUS

SINGLE DATA SOURCE

GEORGE'S DATABASE WILL NEED TO CONSTANTLY UPDATE THE ORDER - SO READ/WRITE SHOULD BOTH BE FAST

FRED

REVENUE ANALYTICS

BATCHES

READ

LONG-RUNNING JOBS

MULTIPLE DATA

GEORGE

ORDER MANAGEMENT

INDIVIDUAL ORDERS

READ /UPDATE

INSTANTANEOUS

SINGLE DATA SOURCE

GEORGE NEEDS ANY UPDATES TO HAPPEN INSTANTANEOUSLY

FRED

REVENUE ANALYTICS

BATCHES

READ

LONG-RUNNING JOBS

GEORGE

ORDER MANAGEMENT

INDIVIDUAL ORDERS

READ /UPDATE

INSTANTANEOUS

IF FRED RUNS HEAVY, LONG-RUNNING JOBS ON THE SAME DATABASE, GEORGE'S UPDATES WILL BE DELAYED

FRED

REVENUE ANALYTICS

BATCHES

READ

LONG-RUNNING JOBS

MULTIPLE DATA SOURCES

GEORGE

ORDER MANAGEMENT

INDIVIDUAL ORDERS

READ /UPDATE

INSTANTANEOUS

SINGLE DATA SOURCE

FRED NEEDS THE DATABASE TO HAVE LOT'S OF INFORMATION FROM MULTIPLE SOURCES

FRED

REVENUE ANALYTICS

BATCHES

READ

LONG-RUNNING JOBS

MULTIPLE DATA SOURCES

GEORGE

ORDER MANAGEMENT

INDIVIDUAL ORDERS

READ /UPDATE

INSTANTANEOUS

SINGLE DATA SOURCE

FRED'S DATABASE WILL BE MUCH, MUCH LARGER IN SCALE

FRED

REVENUE ANALYTICS

BATCHES

READ

LONG-RUNNING JOBS

MULTIPLE DATA SOURCES

GEORGE

ORDER MANAGEMENT

INDIVIDUAL ORDERS

READ /UPDATE

INSTANTANEOUS

SINGLE DATA SOURCE

GEORGE CAN DO WITH A MUCH SMALLER AND COMPACT DATABASE

BATCHES

READ

LONG RUNNING JOBS

MULTIPLE DATA SOURCES

ANALYTICAL PROCESSING

VS

INDIVIDUAL ORDERS

READ/UPDATE

INSTANTANEOUS

SINGLE DATA SOURCE

TRANSACTION PROCESSING

THIS IS A CONUNDRUM THAT'S VERY COMMONLY SEEN

ANALYTICAL PROCESSING

VS

TRANSACTION PROCESSING

IN EARLIER TIMES, THE SCALE OF DATA WAS NOT VERY LARGE

ANALYTICAL PROCESSING

VS

TRANSACTION PROCESSING

BOTH OF THESE TYPES OF NEEDS COULD BE SERVED FROM THE SAME DATABASE

ANALYTICAL PROCESSING

VS

TRANSACTION PROCESSING

AS MORE AND MORE ACTIVITY IS RECORDED ONLINE, THE SCALE OF DATA HAS EXPLODED

BATCHES

READ

LONG-RUNNING JOBS

MULTIPLE DATA SOURCES

ANALYTICAL PROCESSING

VS

INDIVIDUAL ORDERS

READ/UPDATE

INSTANTANEOUS

SINGLE DATA SOURCE

TRANSACTION PROCESSING

THE REQUIREMENTS FOR EACH TYPE OF DATABASE SIGNIFICANTLY DIVERGED

BATCHES

READ

ANALYTICAL

PROCESSING

LONG-RUNNING JOBS

MULTIPLE DATA SOURCES

VS

INDIVIDUAL ORDERS

READ /UPDATE

INSTANTANEOUS

TRANSACTION PROCESSING

TRADITIONAL DATABASES CONTINUE TO BE USED

BATCHES

READ

LONG-RUNNING JOBS

MULTIPLE DATA SOURCES

INDIVIDUAL ORDERS

READ/UPDATE

TRANSACTION

INSTANTANEOUS

SINGLE DATA SOURCE

ANALYTICAL PROCESSING

DATAWAREHOUSE - A DATABASE THAT SPECIFICALLY SERVES ANALYTICAL NEEDS

DATAWAREHOUSE

THERE ARE MANY DATAWAREHOUSE TECHNOLOGIES AVAILABLE TODAY

BATCHES

READ

LONG-RUNNING JOBS

MULTIPLE DATA SOURCES

DATAWAREHOUSE

DATAWAREHOUSE TECHNOLOGIES

VERTICA

TERADATA

ORACLE

IBM

READ

LONG-RUNNING JOBS

MULTIPLE DATA SOURCES

DATAWAREHOUSE

MOST OF THESE TECHNOLOGIES ARE CLOSED-SOURCE

VERTICA

TERADATA

ORACLE

IBM

DATAWAREHOUSE

THEY ARE
PROPRIETARY AND PAID
SOFTWARES

VERTICA

TERADATA

ORACLE

IBM

HIVE IS AN OPEN-SOURCE DATAWAREHOUSE

HIVE

HIVE WAS DEVELOPED BY THE OPEN-SOURCE POWERHOUSE

APACHE

HIVE

HIVE IS PART OF A LARGER ECOSYSTEM OF OPEN-SOURCE DISTRIBUTED COMPUTING APPLICATIONS

HIVE

THE APACHE DISTRIBUTED COMPUTING ECOSYSTEM IS BUILT AROUND

HADOOP

HADOOP

is a distributed computing framework
developed and maintained by

THE APACHE SOFTWARE FOUNDATION

written in Java

HADOOP

HDFS

A file system to manage the storage of data

HADOOP

HDFS

MapReduce

A framework to process data across multiple servers

HADOOP

HDFS

YARN

A framework to run the data processing task

MapReduce

A framework to define a data processing task

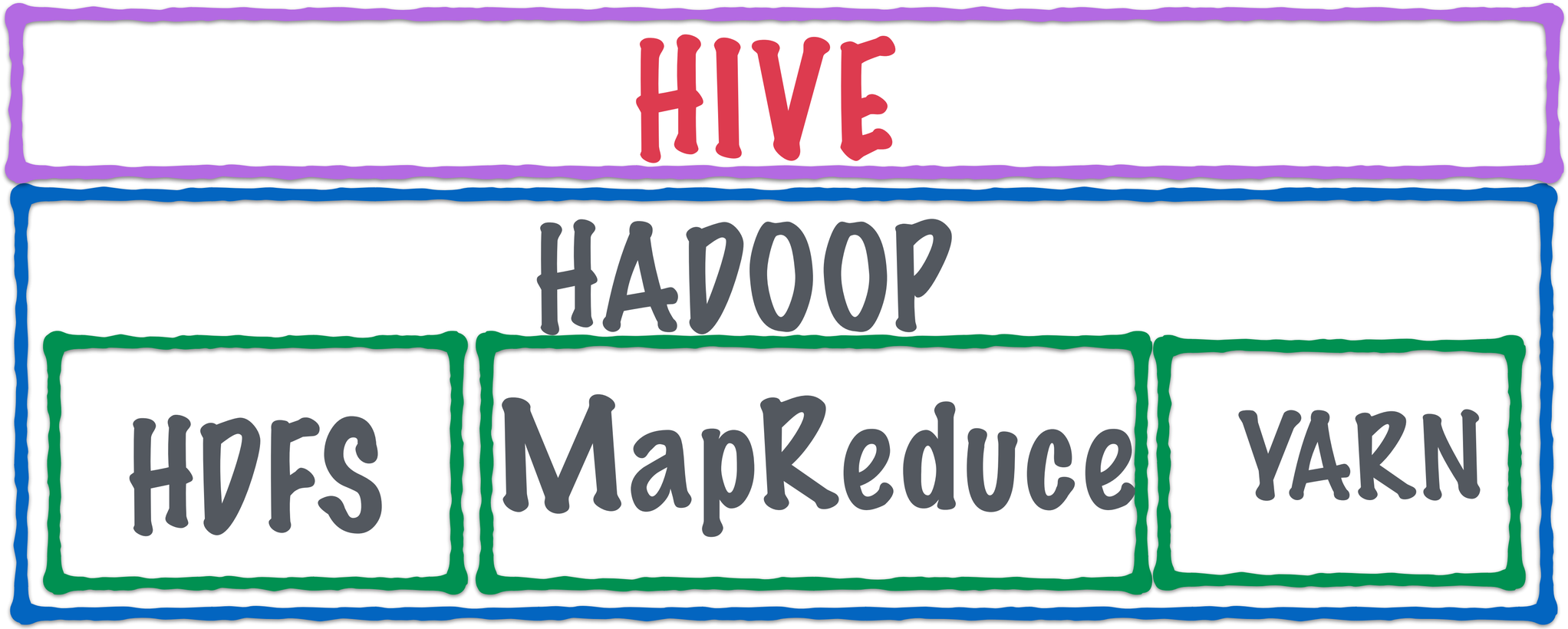
HADOOP

HDFS

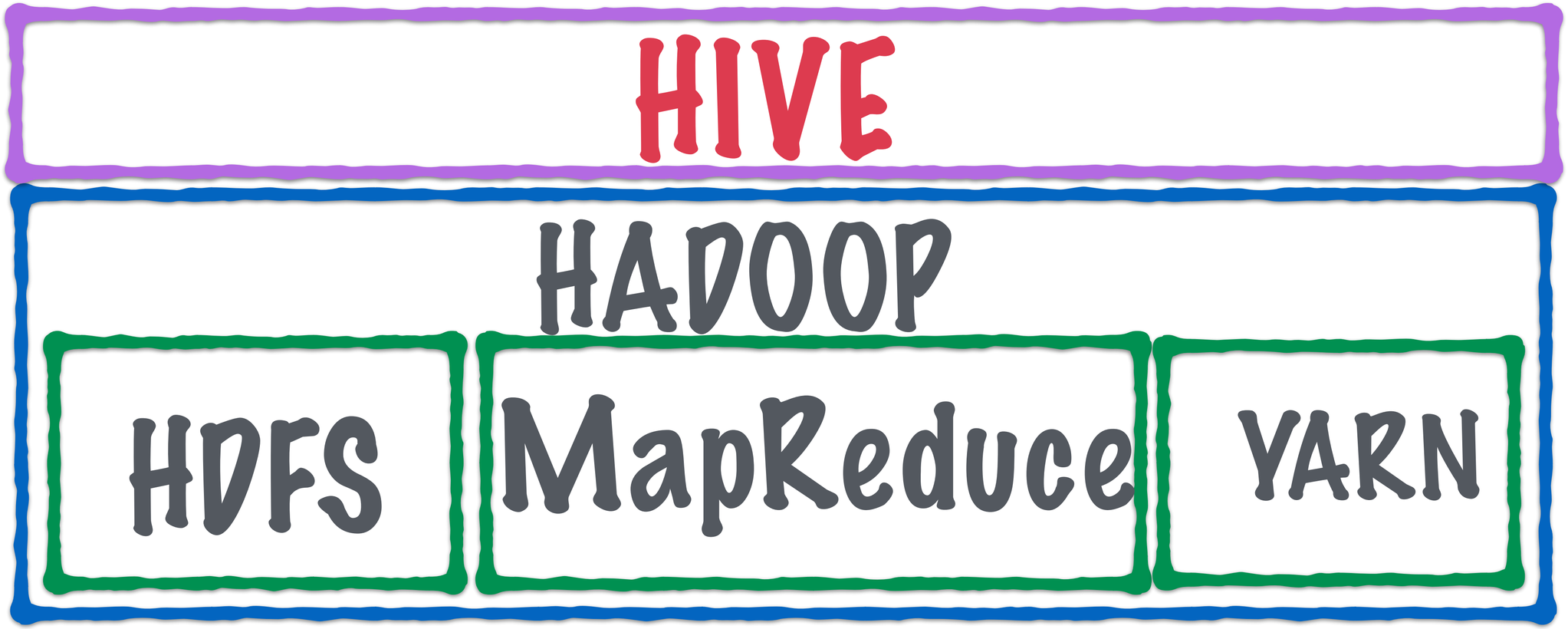
YARN

MapReduce

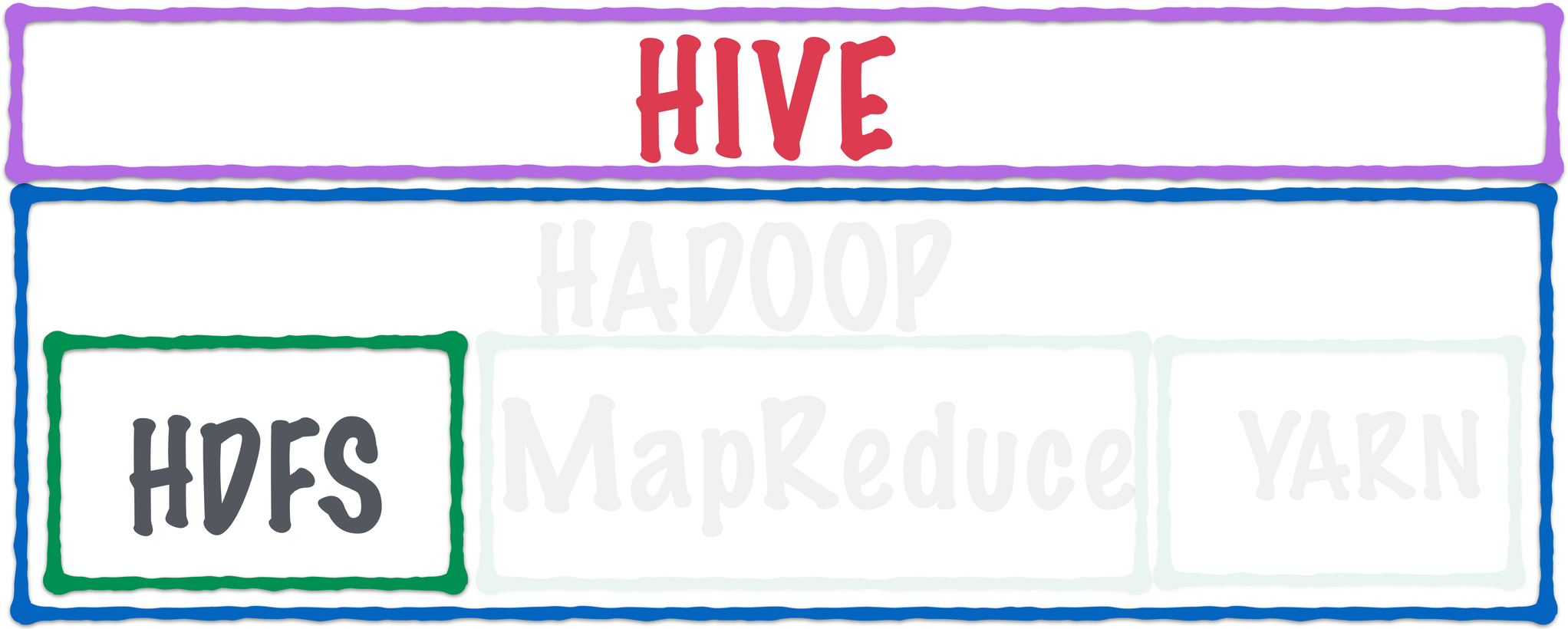
We'll go through each of these blocks a little later



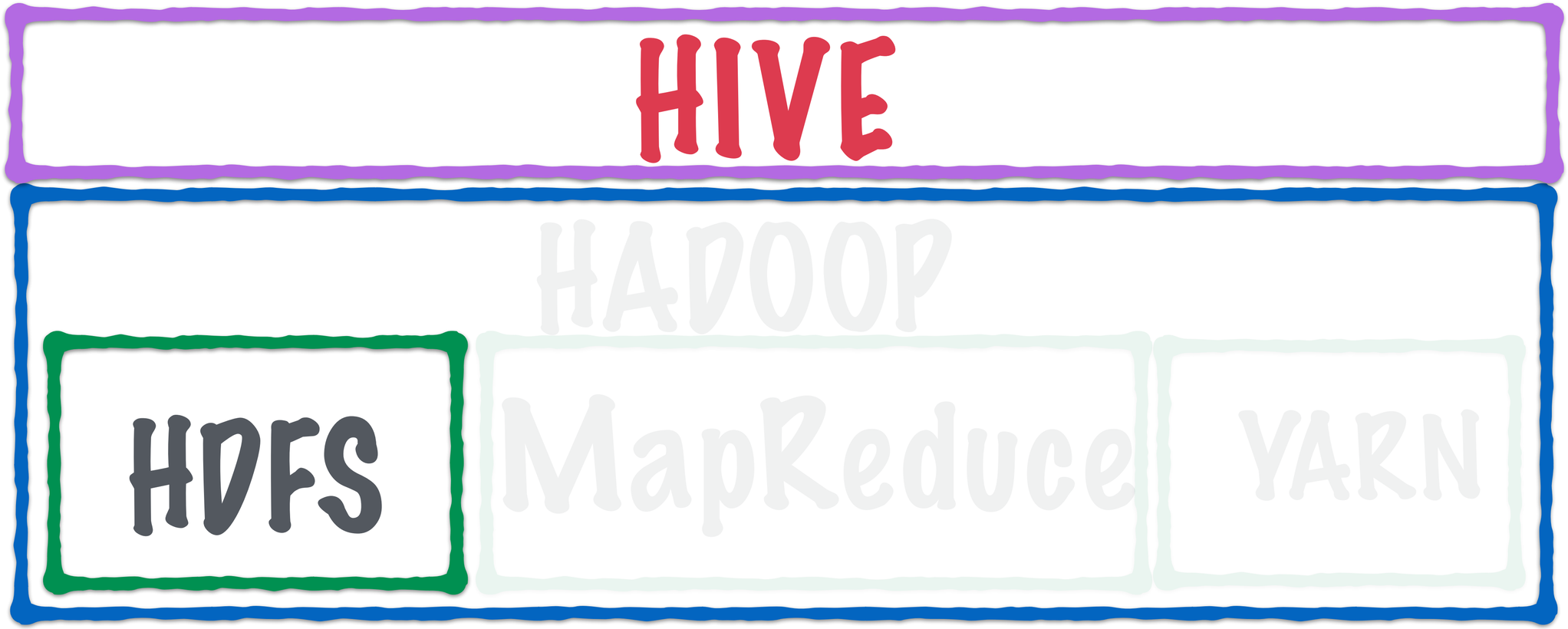
For now, let's quickly understand
how Hive uses these



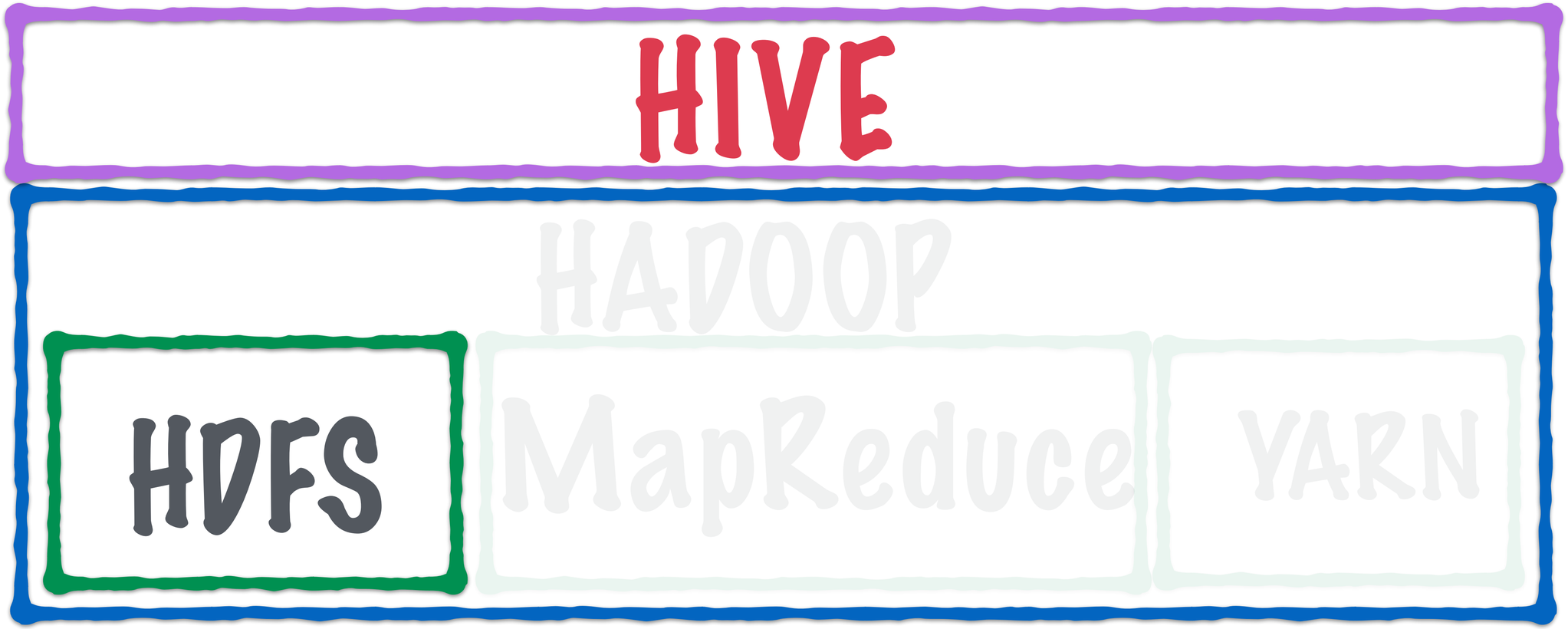
HIVE IS A DATAWAREHOUSE
BUILT ON TOP OF HADOOP



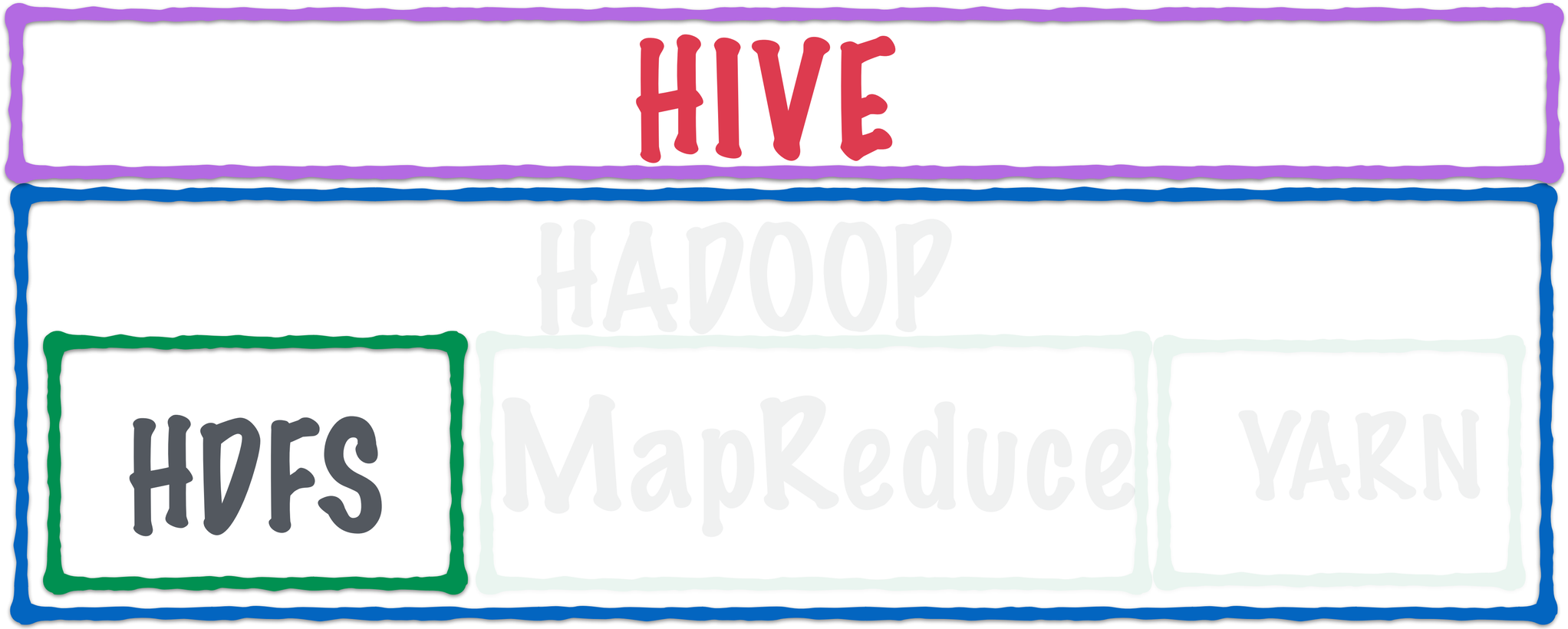
HIVE STORES IT'S DATA
AS FILES IN HDFS



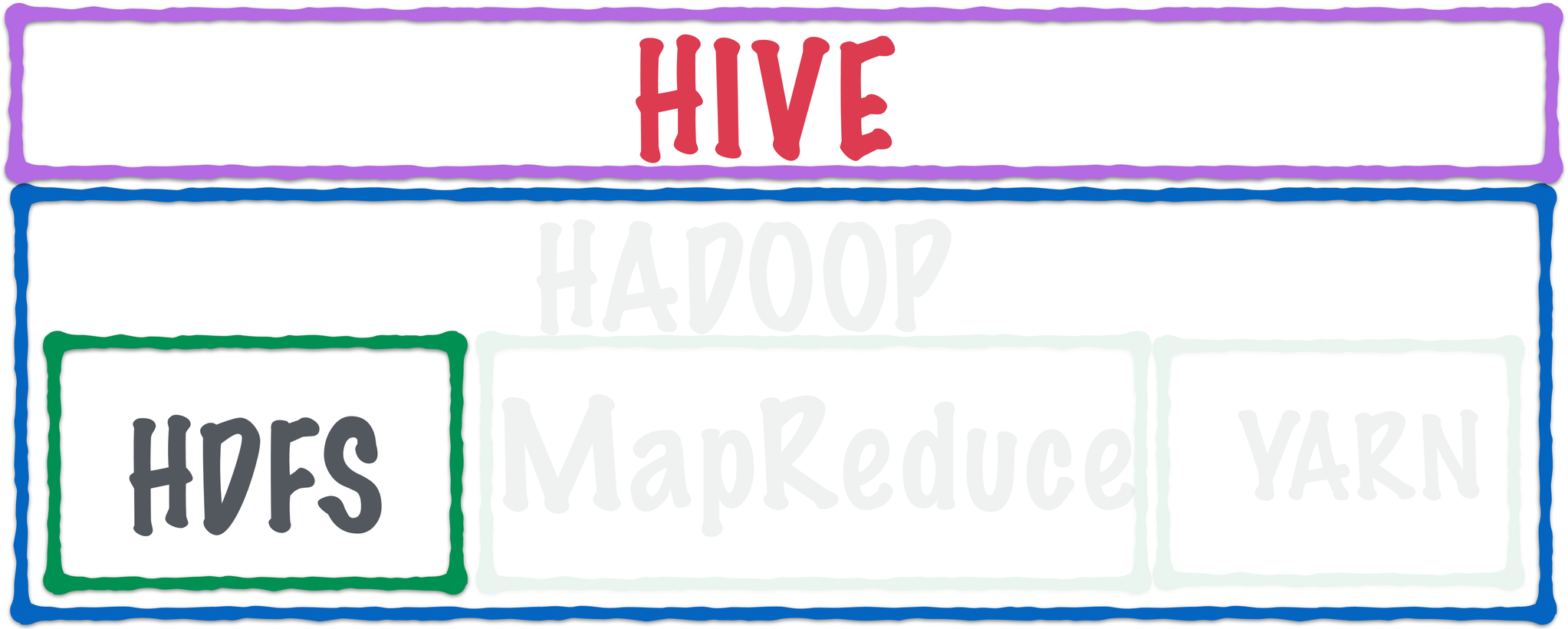
HDFS stores the files by distributing them across multiple machines



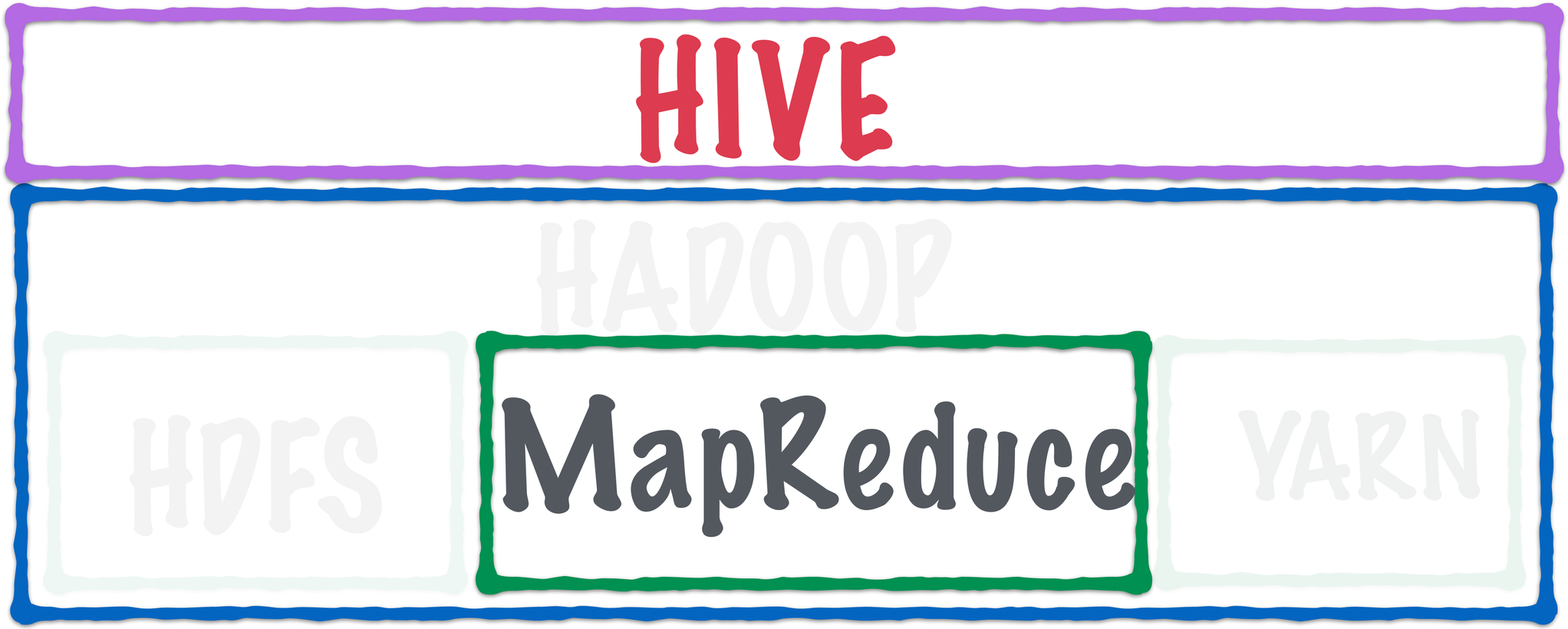
WHY DOES THIS HELP?



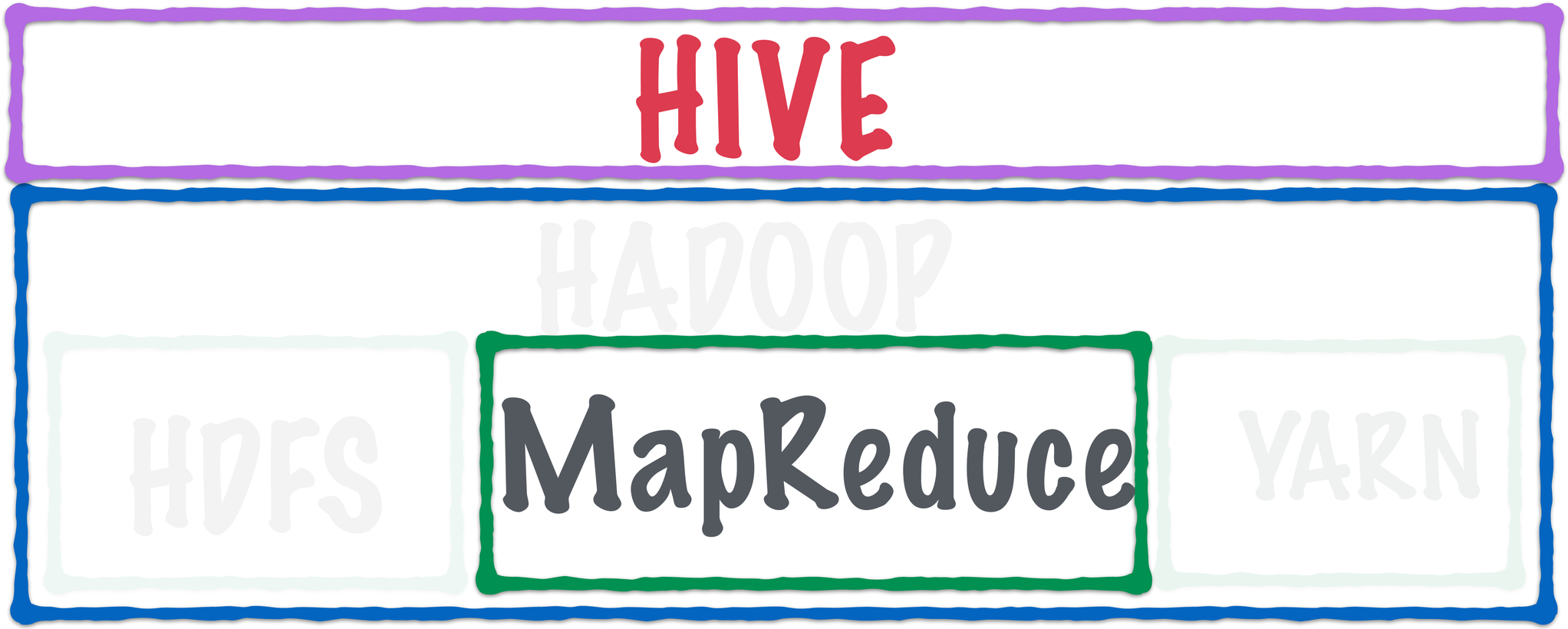
Because Hadoop can parallelize any processing tasks on that data



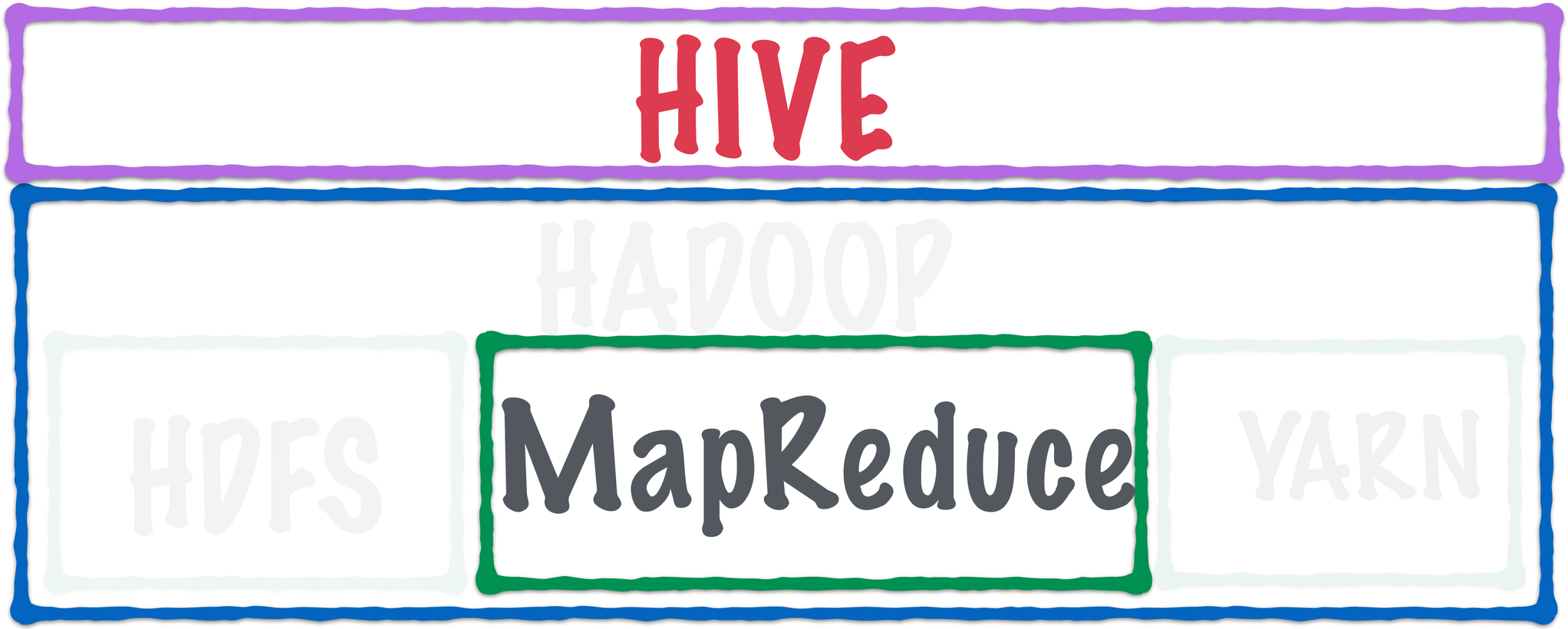
This makes the processing much faster than if you had stored it on a single machine



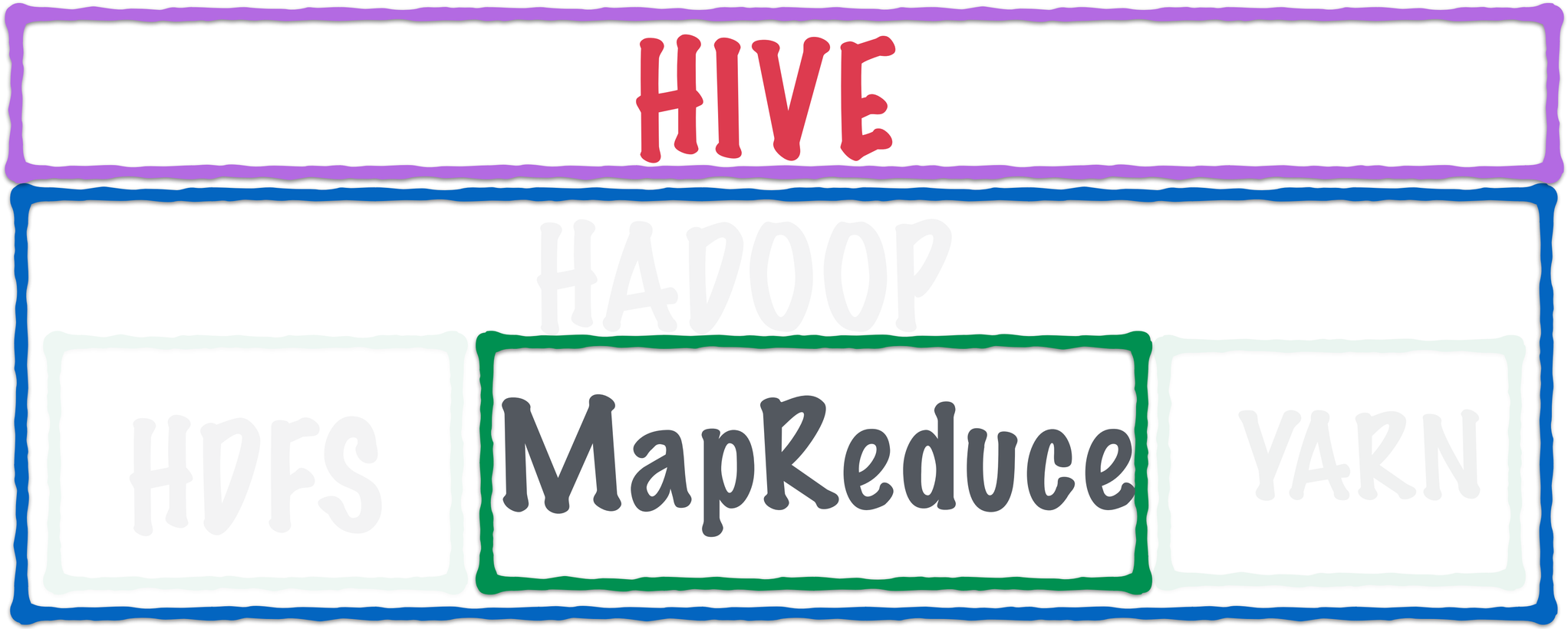
All processing tasks in Hadoop are run using MapReduce tasks



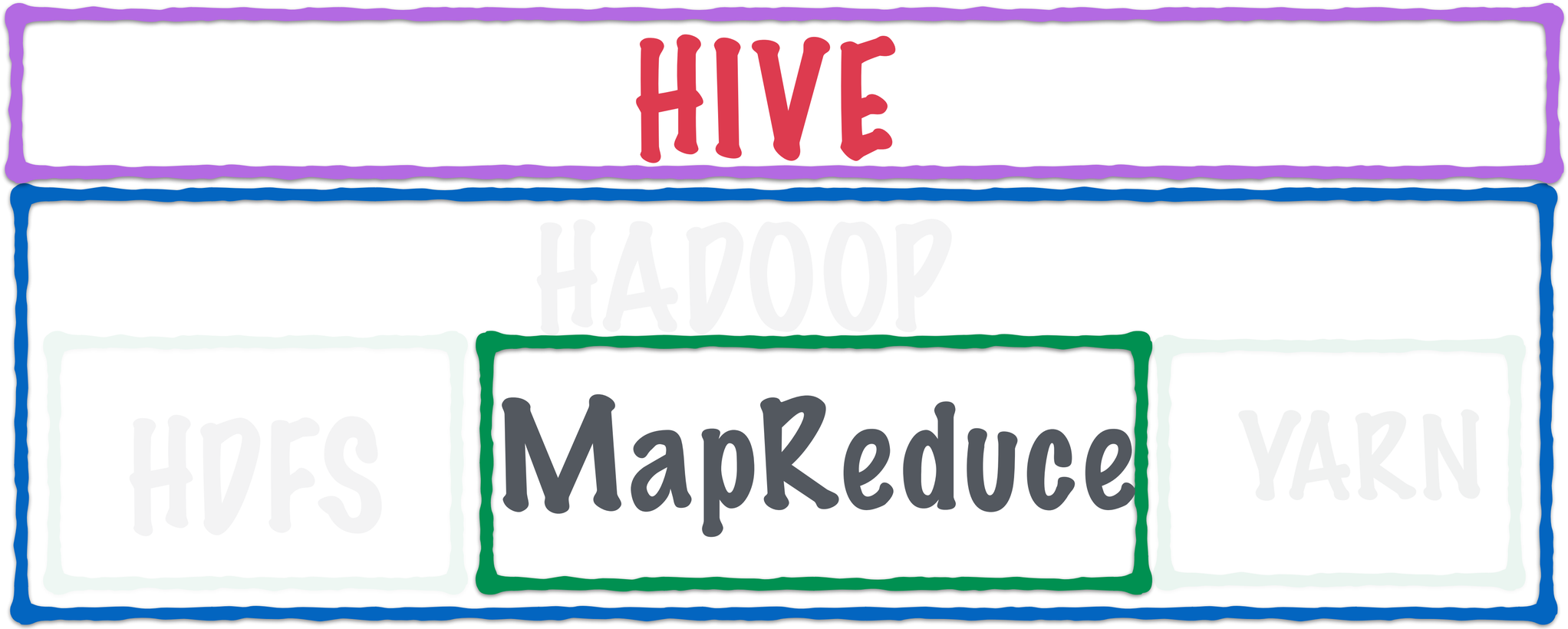
MapReduce tasks are usually written using a Java Framework



Writing these MapReduce tasks can be pretty daunting



Traditional databases/closed-source datawarehouses normally use SQL



SQL = Structured Query Language

SQL = Structured Query Language

SQL is really much easier to use and understand :)

SQL = Structured Query Language

It's widely used by analysts and programmers to work with databases/data warehouses

SQL = Structured Query Language

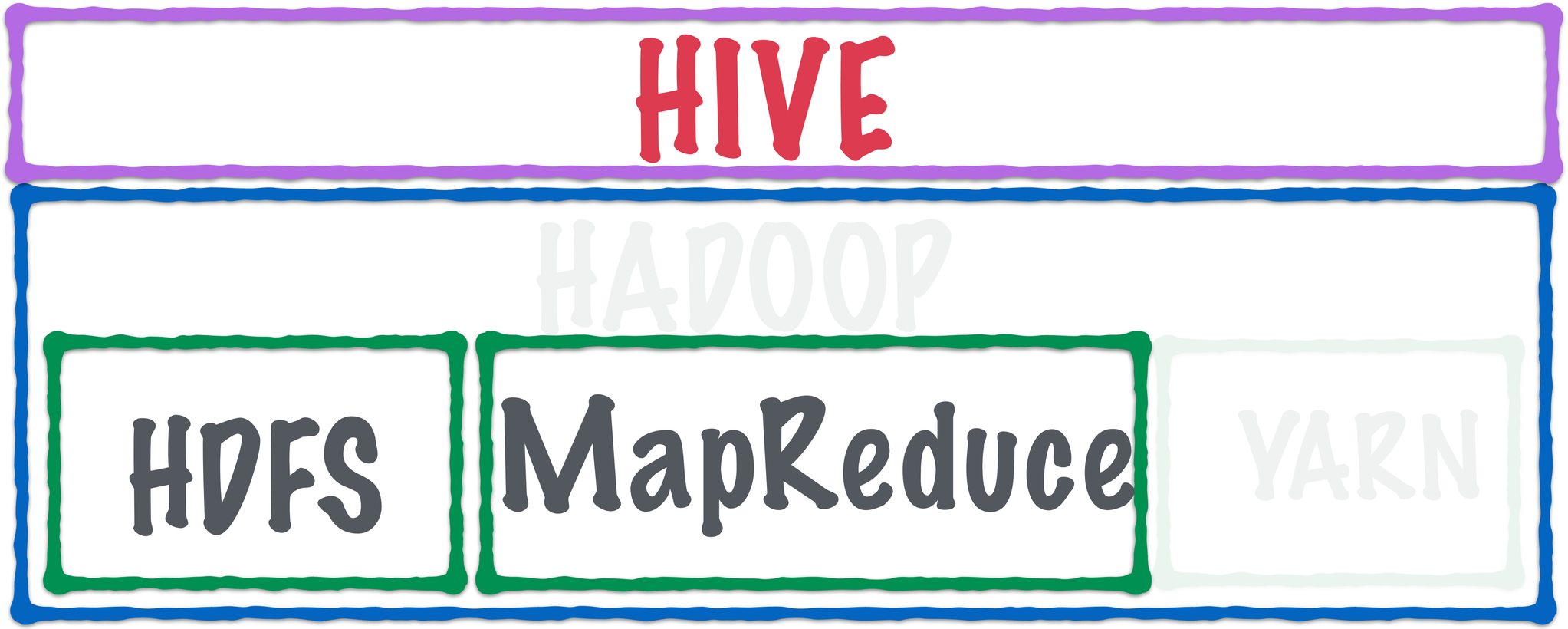
SQL has a few easy to understand constructs

Select, group by, join etc

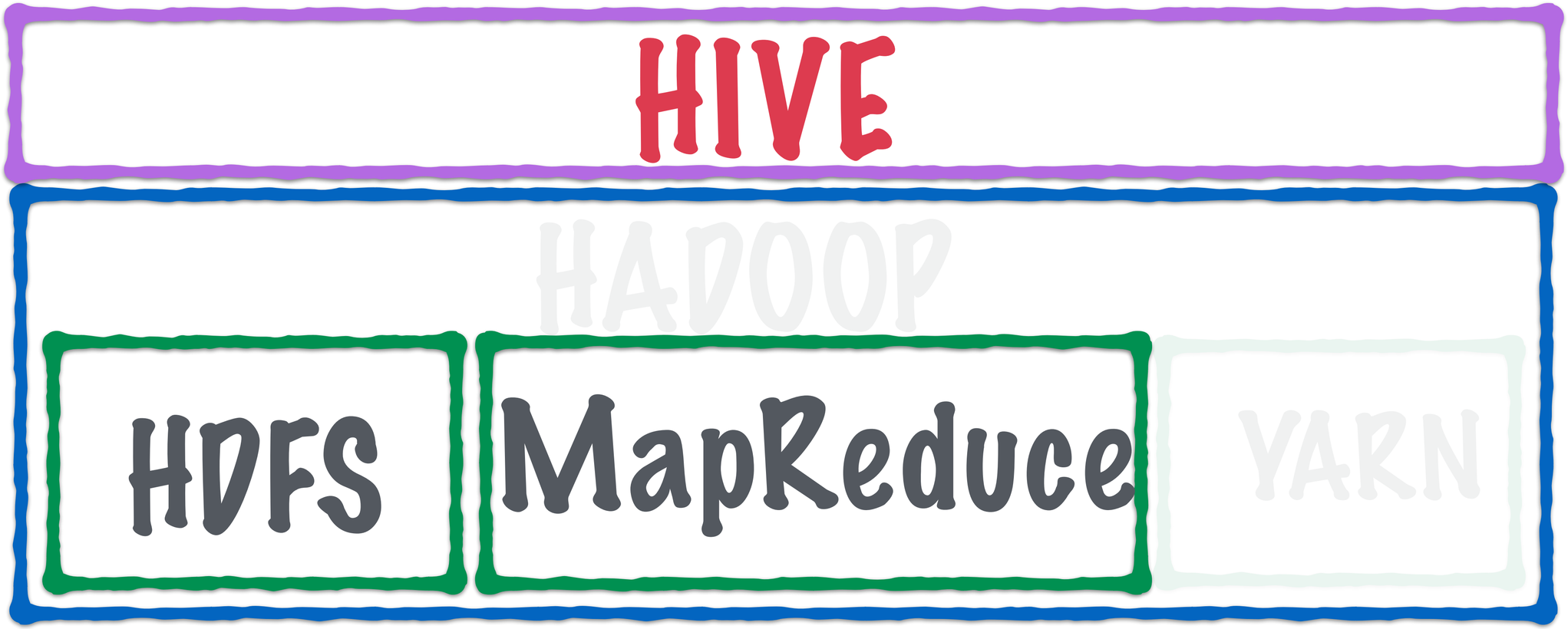
SQL = Structured Query Language

Most data processing tasks are defined using a combination of these constructs

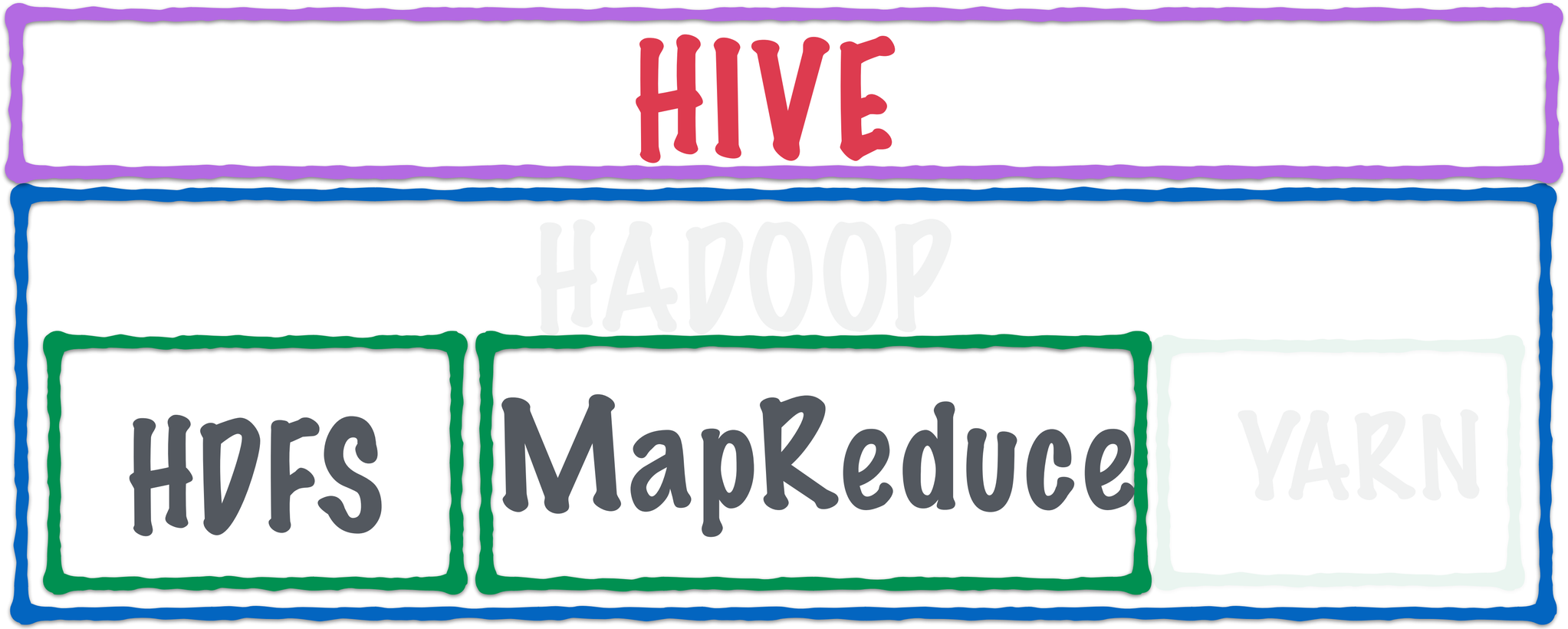
Select, group by, join etc



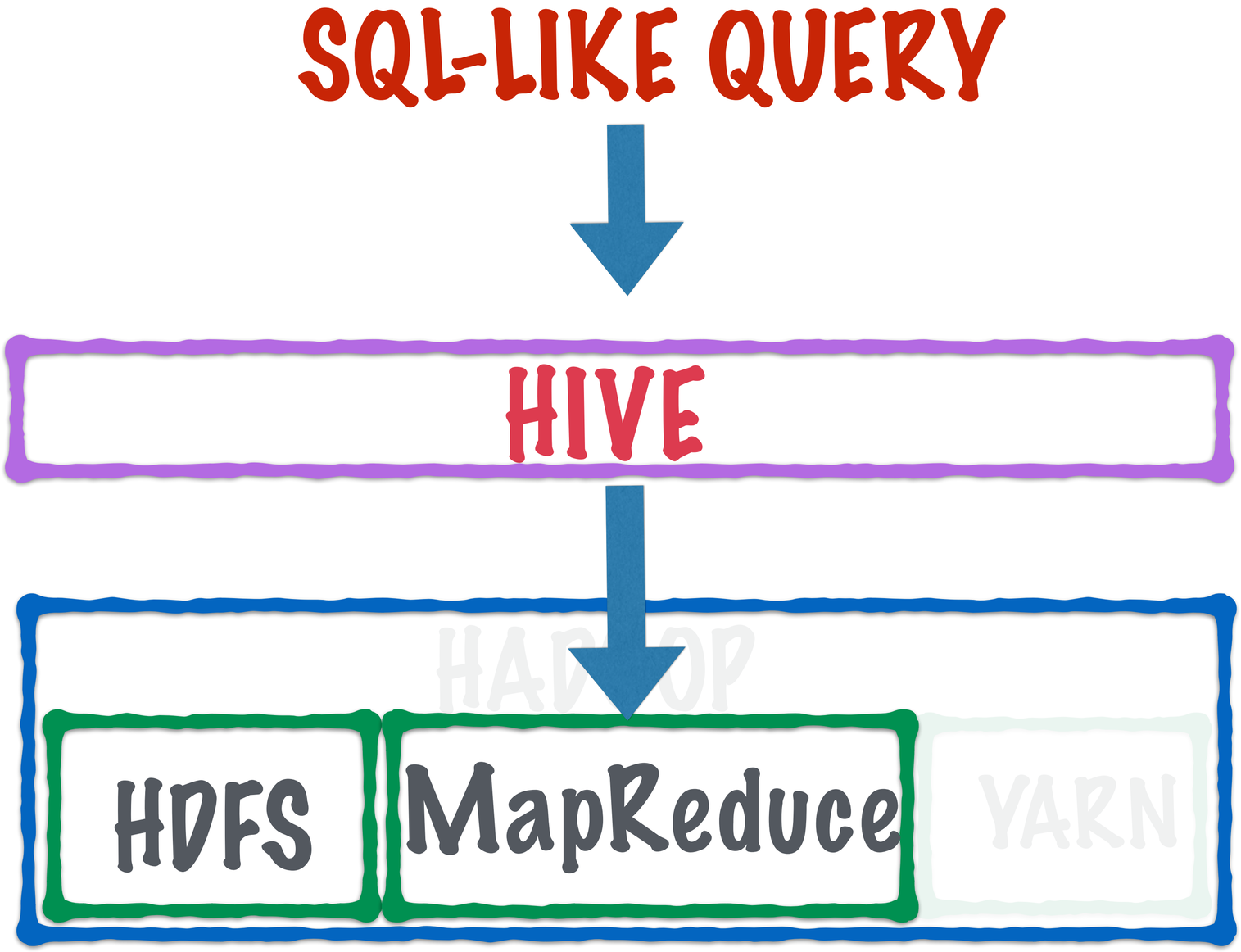
HIVE PROVIDES AN SQL LIKE INTERFACE TO DATA IN HDFS



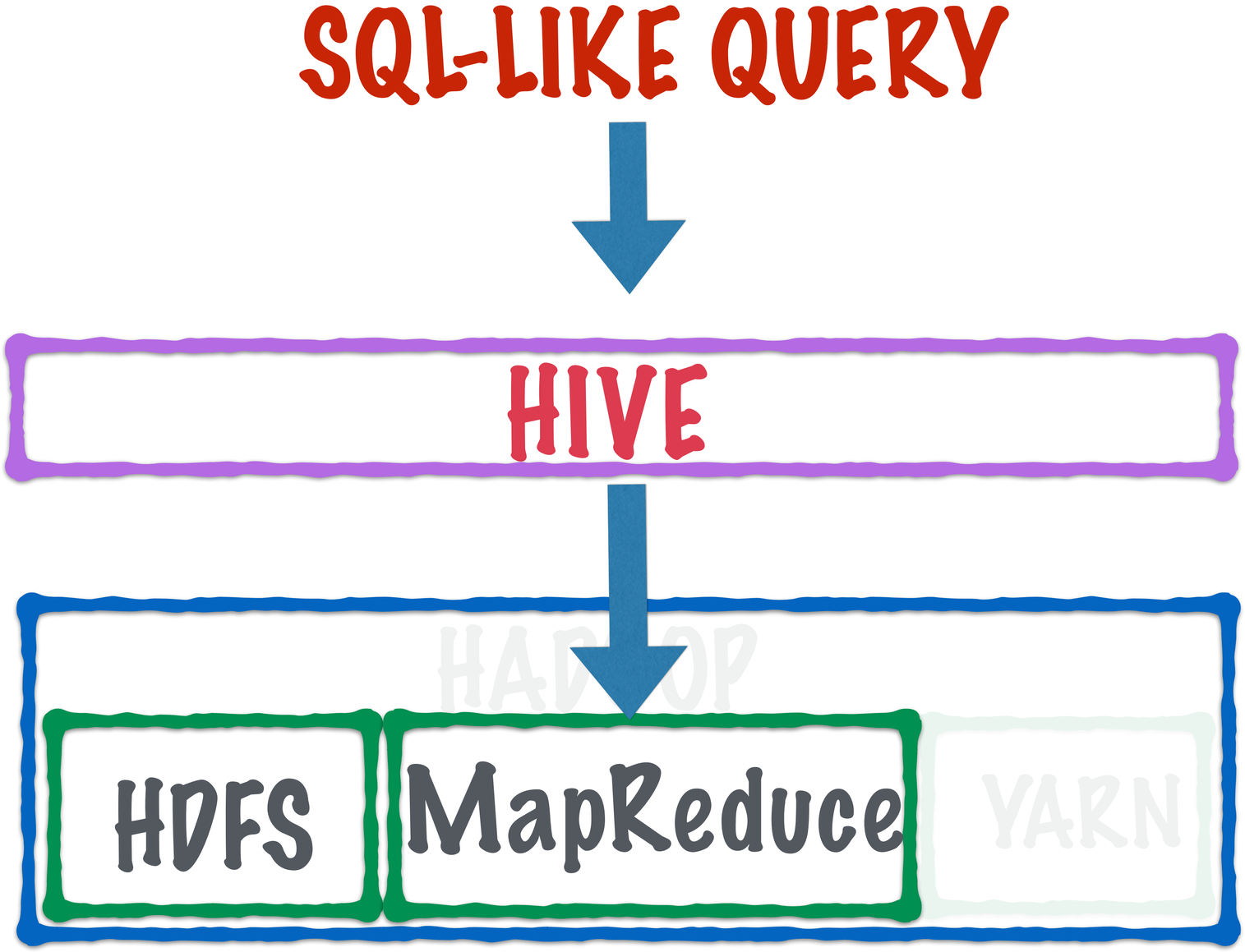
THE FILES IN HDFS ARE EXPOSED TO THE USER IN THE FORM OF TABLES



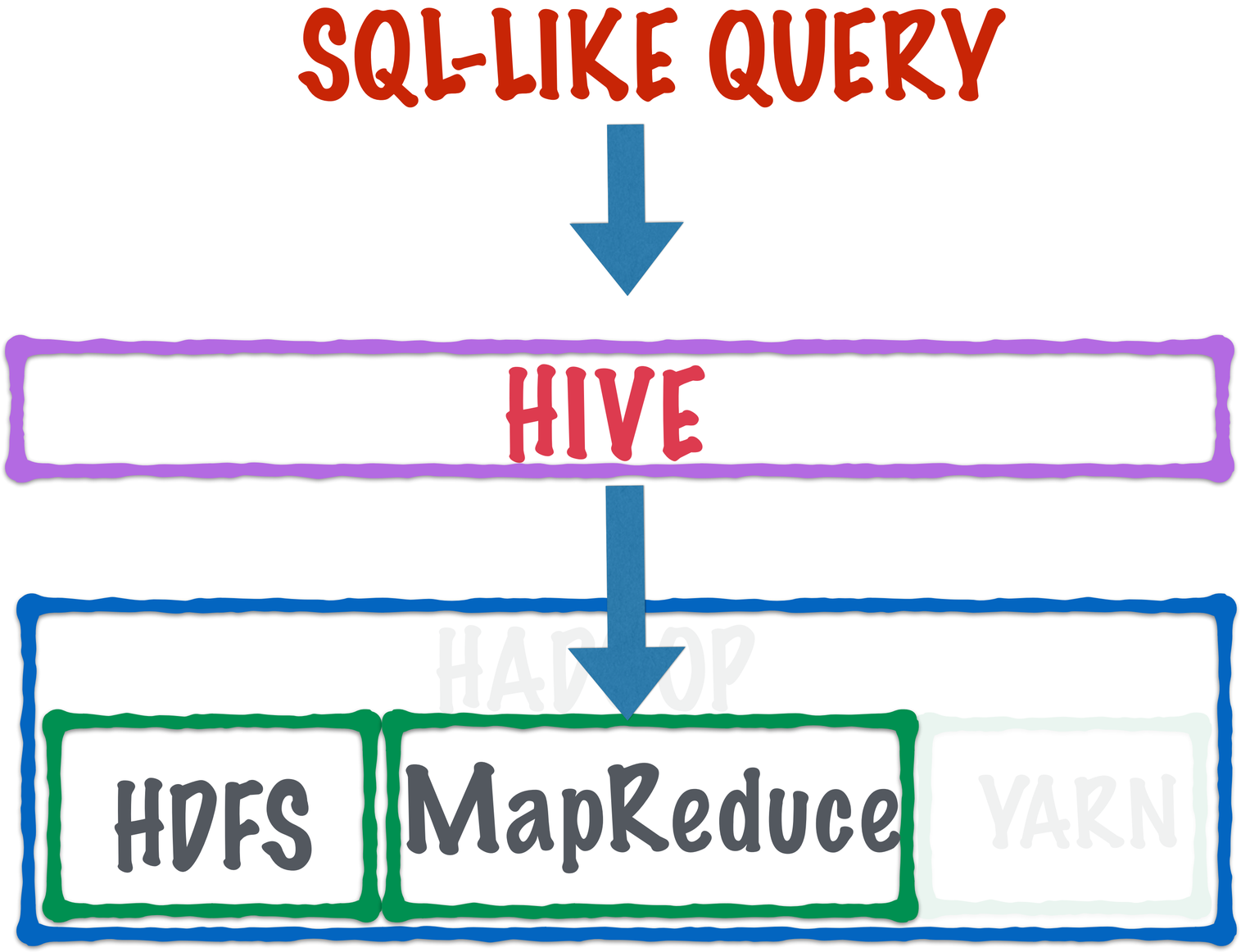
THE USER CAN WRITE SQL-LIKE QUERIES TO WORK WITH THESE TABLES



HIVE WILL TRANSLATE THE QUERY INTO 1/MORE MAPREDUCE TASKS

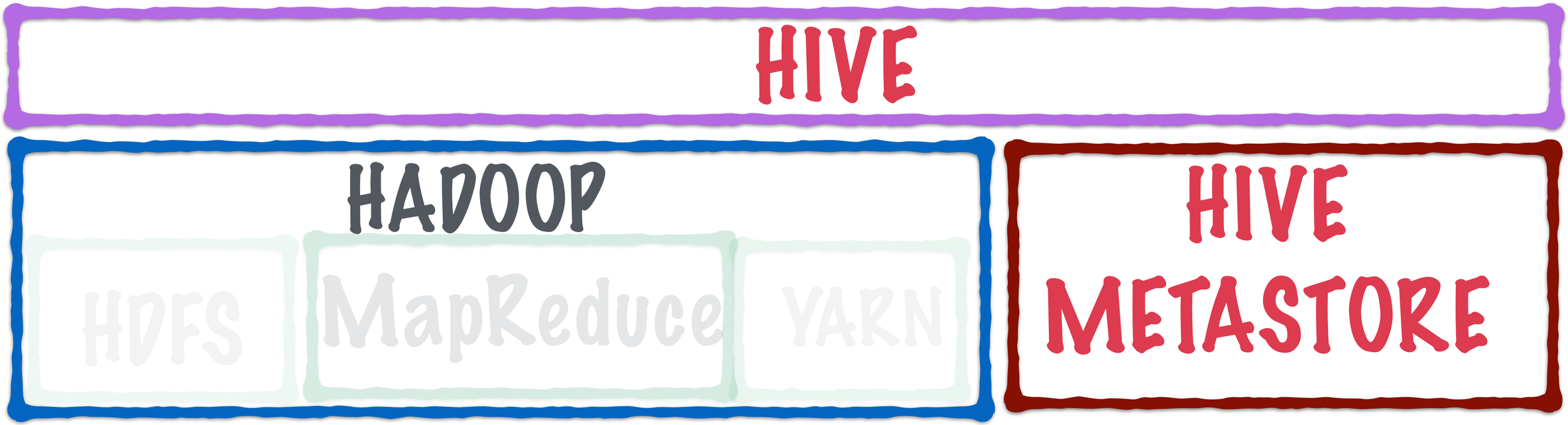


THE MAPREDUCE TASKS WILL PROCESS THE DATA IN HDFS AND RETURN ANY RESULTS TO HIVE

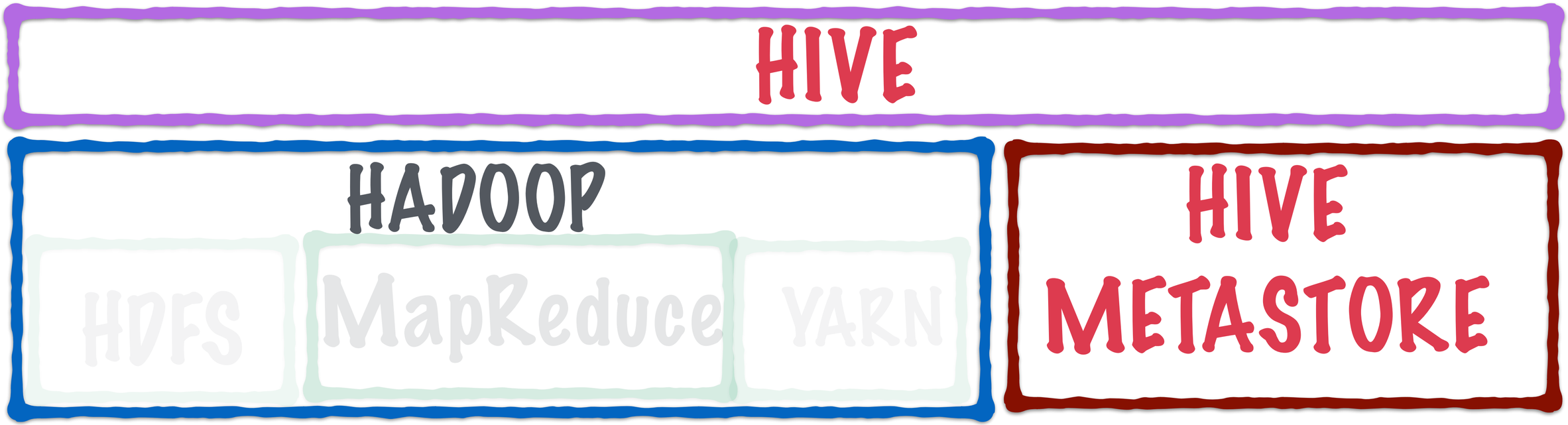


THE QUERIES ARE WRITTEN IN A SQL LIKE LANGUAGE CALLED HIVEQL

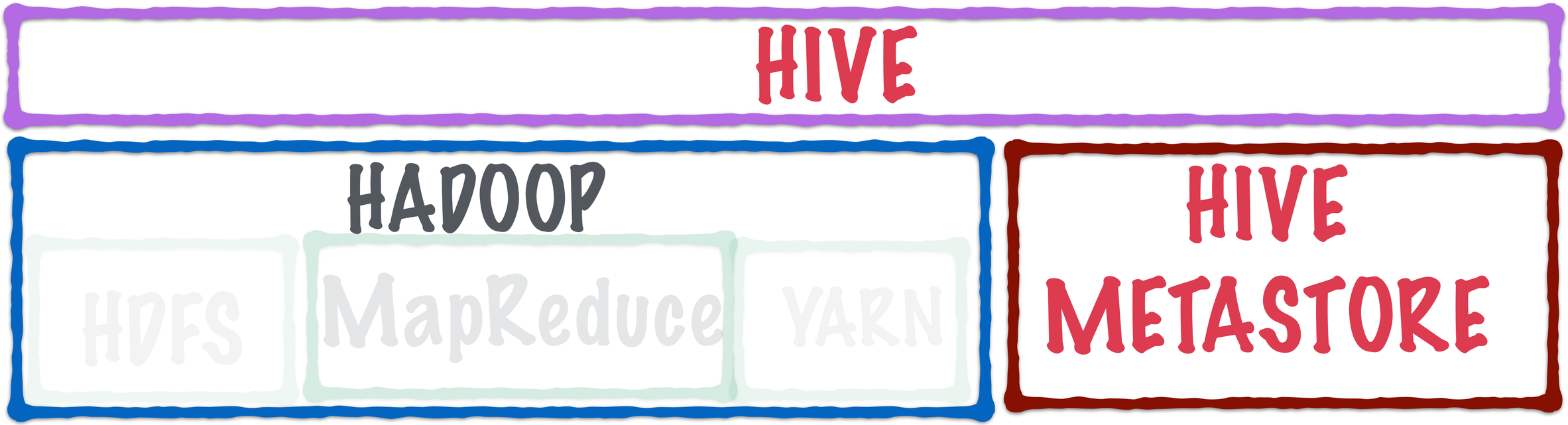
HIVE HAS 1 MORE IMPORTANT COMPONENT ALONG WITH HADOOP



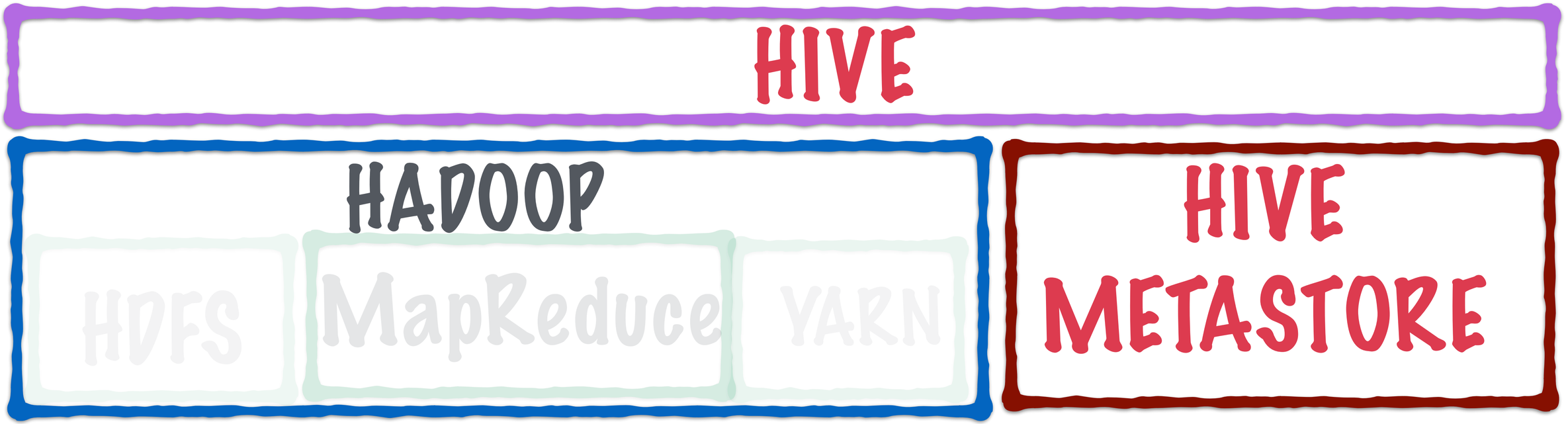
THE METASTORE IS A DATABASE



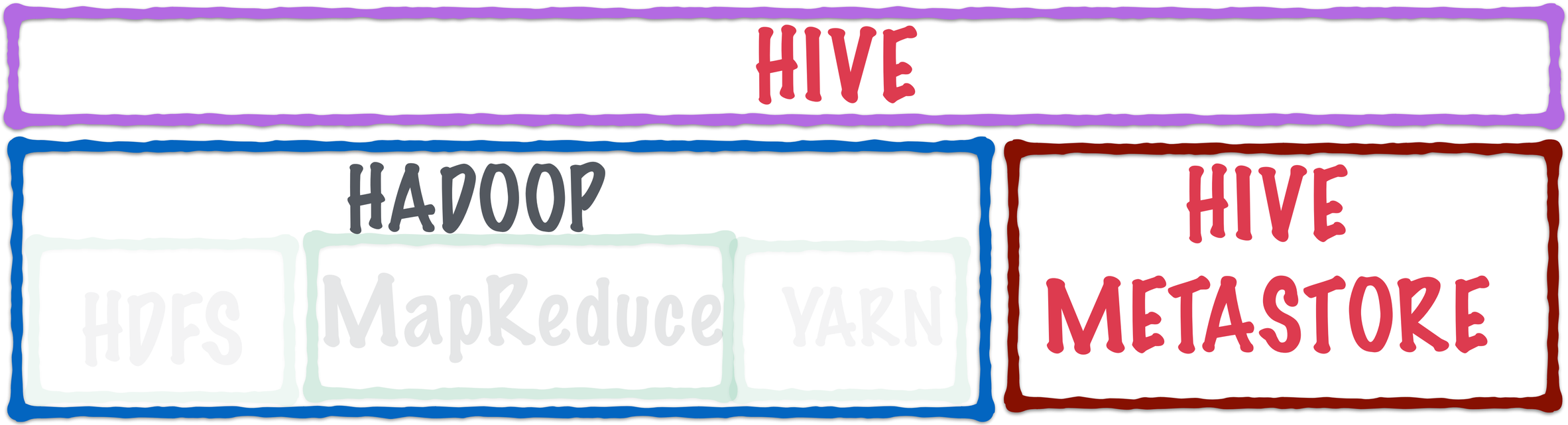
THE HIVE METASTORE STORES ALL THE METADATA FOR THE TABLES IN HIVE



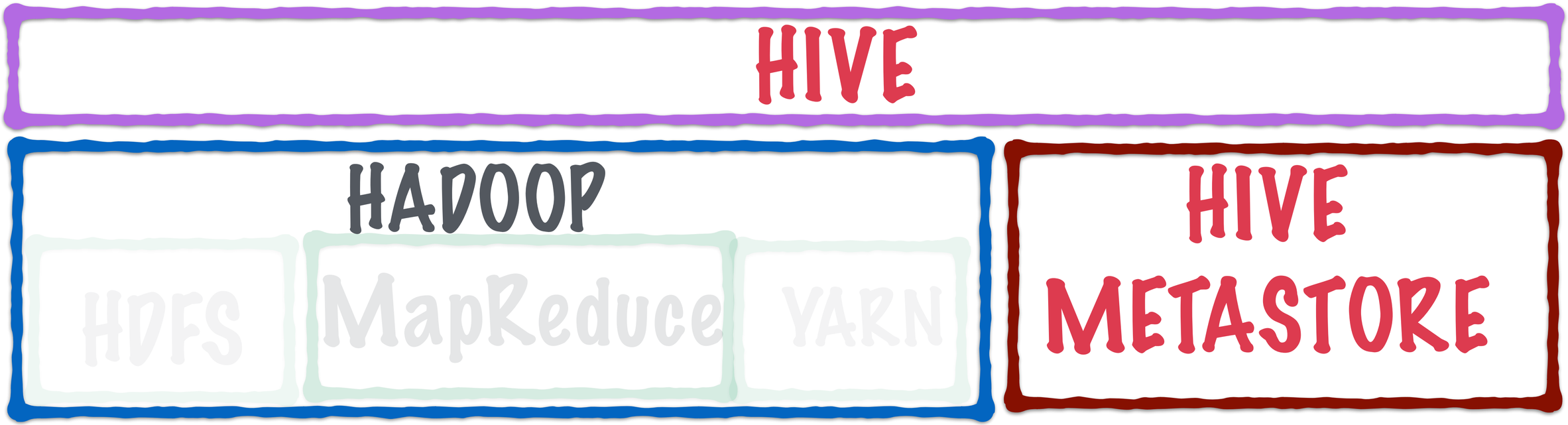
THE HIVE DATA IN HDFS IS IN THE FORM OF FILES AND DIRECTORIES



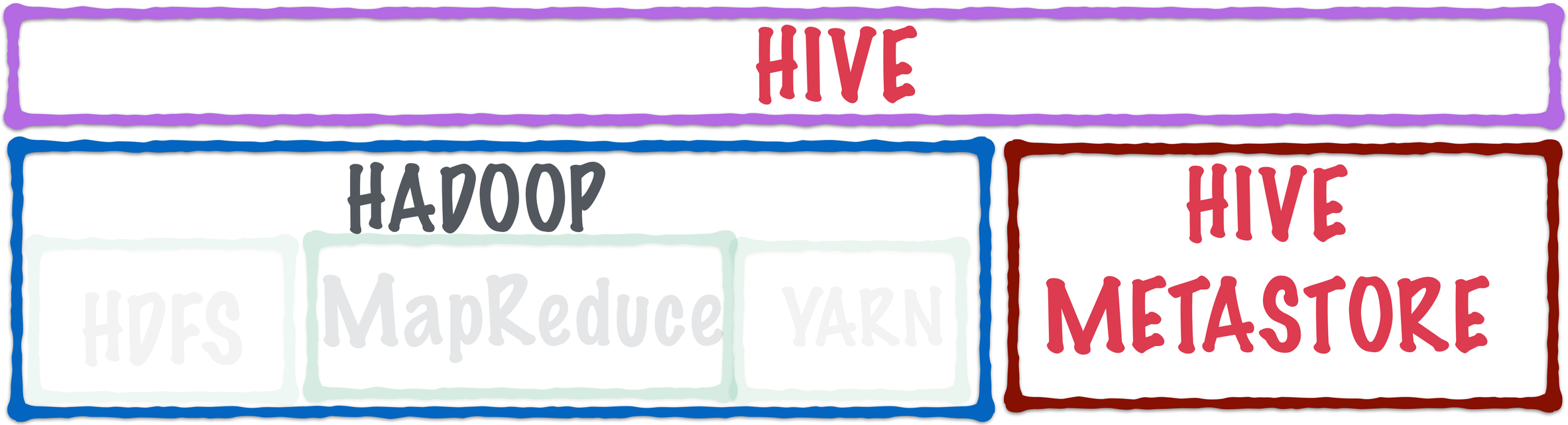
THE METASTORE MAPS THESE DIRECTORIES AND FILES TO HIVE TABLES



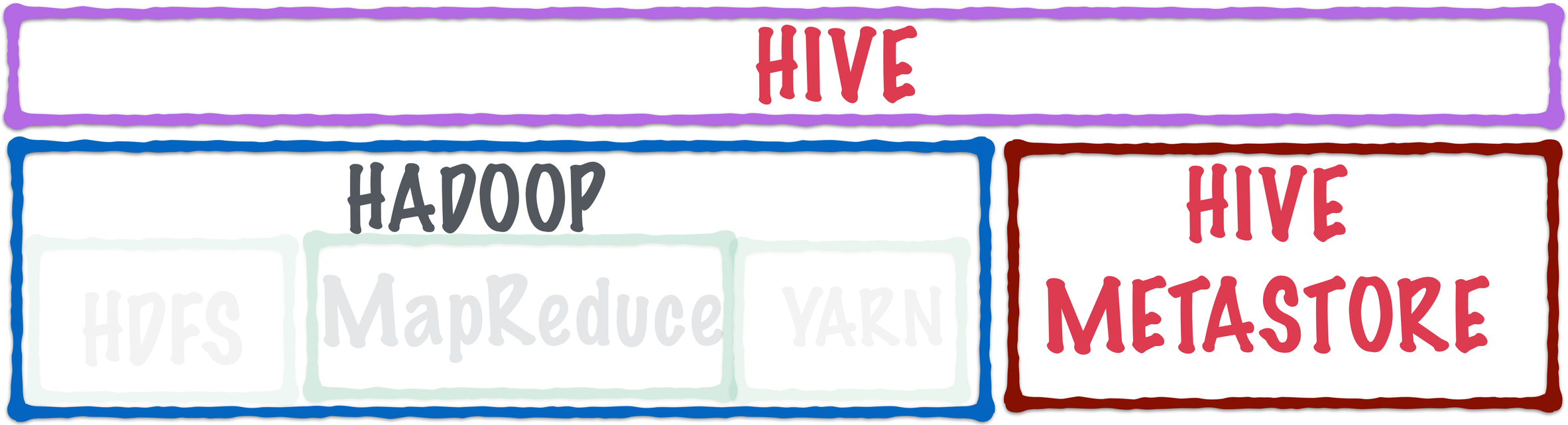
IT HOLDS THE TABLE DEFINITIONS AND SCHEMA



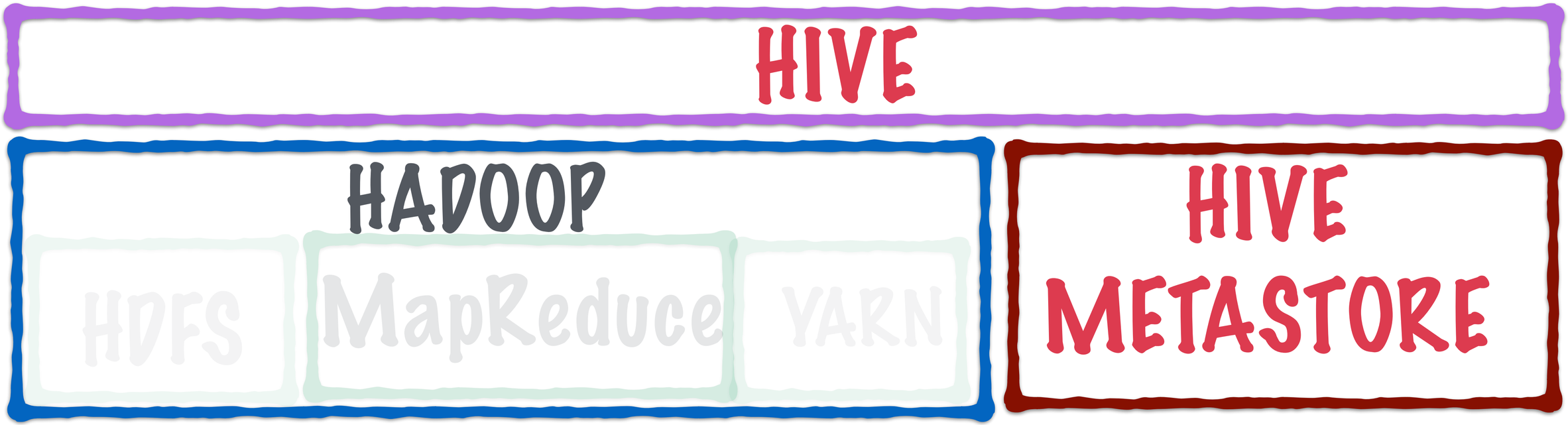
THE SCHEMA CONSISTS OF COLUMN DEFINITIONS AND DATA TYPES



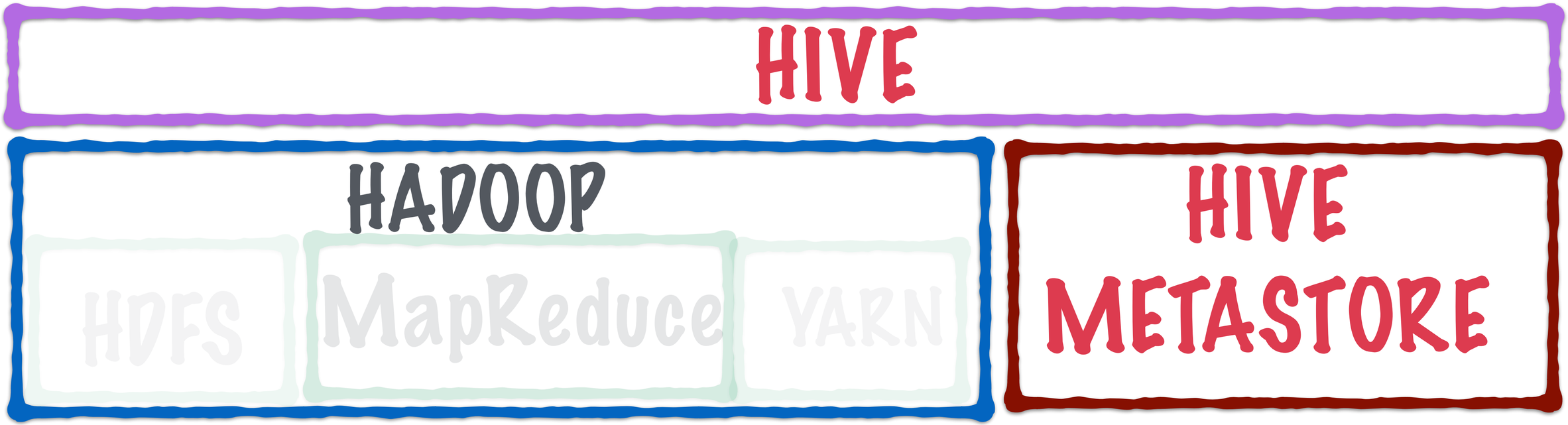
THE METADATA ALSO TELLS HIVE HOW TO READ A FILE AND CONVERT INTO A TABLE / COLUMN REPRESENTATION



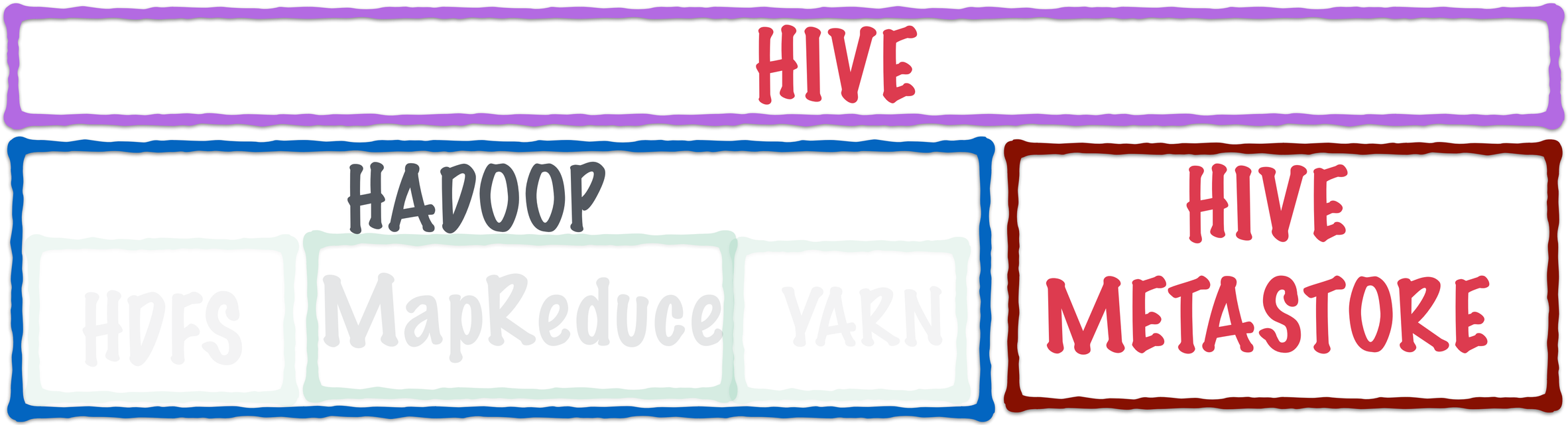
FOR EX: IF THE FILE IS A CSV FILE, THE METASTORE WILL KNOW HOW TO PARSE THE ROWS IN THE CSV FILE INTO COLUMNS



HIVE ALLOWS THE USER TO CONFIGURE THE HIVE METASTORE



WE CAN USE ANY TRADITIONAL RDBMS TECHNOLOGY



THE DEFAULT IS AN OPEN-SOURCE DATABASE BY APACHE CALLED DERBY