数据应用学院

hive

noun | \ˈhīv\

: a nest for bees

: the bees living in a hive

: a place filled with busy activity

Disclaimer

- All data and information in this presentation were from public resource
- This is a vendor-independent talk that expresses teacher own opinions
- Copyright 2017@ Data Application Institute
- **Please DO NOT record video**

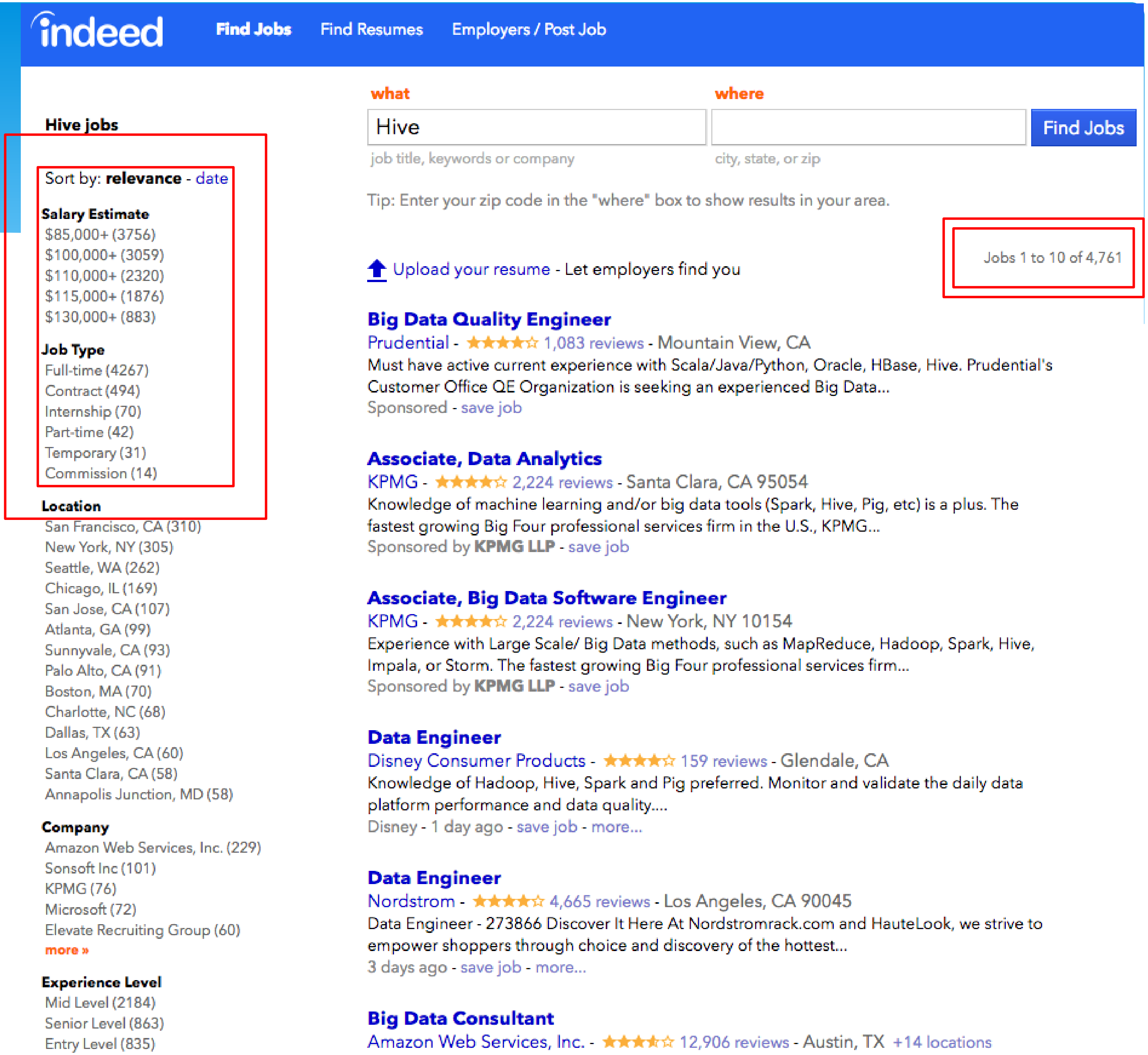
Agenda

- Why we learn Hive
- Hive Introduction
- How Hive works
- HQL programming
- Demo
- Best Practices for High Performance
- Optimization Hive Usage
- Most asked Interview questions

Data Application Lab

Why we need to learn Hive?

- Why?
- Make the world a better place?
- Let’s get real !!



Hive jobs

what: Hive

job title, keywords or company

where:

city, state, or zip

Tip: Enter your zip code in the "where" box to show results in your area.

Upload your resume - Let employers find you

Jobs 1 to 10 of 4,761

Sort by: **relevance** - date

Salary Estimate
$85,000+ (3756)
$100,000+ (3059)
$110,000+ (2320)
$115,000+ (1876)
$130,000+ (883)

Job Type
Full-time (4267)
Contract (494)
Internship (70)
Part-time (42)
Temporary (31)
Commission (14)

Location
San Francisco, CA (310)
New York, NY (305)
Seattle, WA (262)
Chicago, IL (169)
San Jose, CA (107)
Atlanta, GA (99)
Sunnyvale, CA (93)
Palo Alto, CA (91)
Boston, MA (70)
Charlotte, NC (68)
Dallas, TX (63)
Los Angeles, CA (60)
Santa Clara, CA (58)
Annapolis Junction, MD (58)

Company
Amazon Web Services, Inc. (229)
Sonsoft Inc (101)
KPMG (76)
Microsoft (72)
Elevate Recruiting Group (60)
more »

Experience Level
Mid Level (2184)
Senior Level (863)
Entry Level (835)

Big Data Quality Engineer
Prudential - ★★★★☆ 1,083 reviews - Mountain View, CA
Must have active current experience with Scala/Java/Python, Oracle, HBase, Hive. Prudential's Customer Office QE Organization is seeking an experienced Big Data...
Sponsored - save job

Associate, Data Analytics
KPMG - ★★★★☆ 2,224 reviews - Santa Clara, CA 95054
Knowledge of machine learning and/or big data tools (Spark, Hive, Pig, etc) is a plus. The fastest growing Big Four professional services firm in the U.S., KPMG...
Sponsored by **KPMG LLP** - save job

Associate, Big Data Software Engineer
KPMG - ★★★★☆ 2,224 reviews - New York, NY 10154
Experience with Large Scale/ Big Data methods, such as MapReduce, Hadoop, Spark, Hive, Impala, or Storm. The fastest growing Big Four professional services firm...
Sponsored by **KPMG LLP** - save job

Data Engineer
Disney Consumer Products - ★★★★☆ 159 reviews - Glendale, CA
Knowledge of Hadoop, Hive, Spark and Pig preferred. Monitor and validate the daily data platform performance and data quality....
Disney - 1 day ago - save job - more...

Data Engineer
Nordstrom - ★★★★☆ 4,665 reviews - Los Angeles, CA 90045
Data Engineer - 273866 Discover It Here At Nordstromrack.com and HauteLook, we strive to empower shoppers through choice and discovery of the hottest...
3 days ago - save job - more...

Big Data Consultant
Amazon Web Services, Inc. - ★★★★☆ 12,906 reviews - Austin, TX +14 locations

Let’s learn Hive!

- $$$$

- In many use cases, data must be
  - A. loaded into file system
  - B. some schema must be applied
  - C. transformed
  - D. analyzed
  - E. and then visualized

- Hive is a mostly commonly used tool for Big Data.

- A must-have skill for DS, DE, BA and BI.

Apache Hive Origin

- Started at Facebook in 2008 to manage lots of data
- The data was stored in Oracle database every night
- ETL was performed on data
- Data growth was exponential
  - By 2006 1 TB / Day
  - By 2010 10 TB/Day…
- And there was a need to find some way to manage the data "effectively"
- Convert SQL query into a series of MR jobs

Data Application Lab

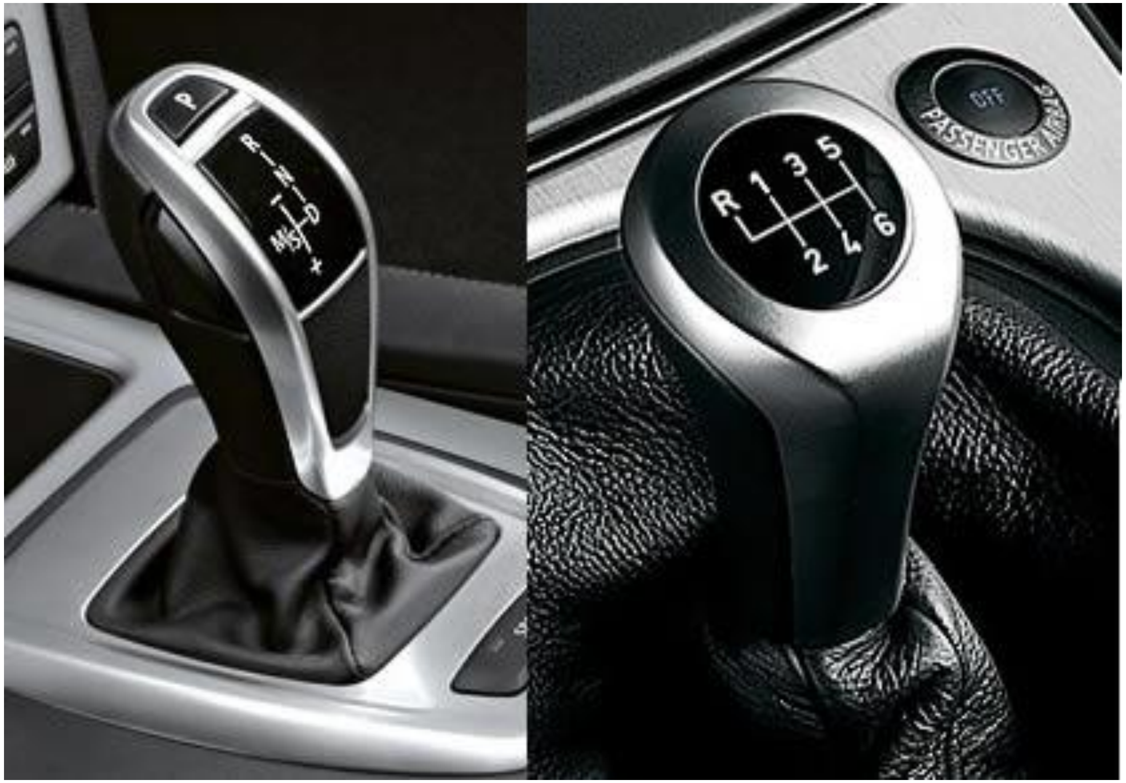
Motivation

- MapReduce has limitations
  - Have to use M/R model
  - Not Reusable
  - Long development type / overhead
  - For complex jobs:
    - Multiple M/R stages
- Bright side:
  - MapReduce is scalable
  - SQL has huge user base
  - SQL is easy to code
- => Solution: Combine MapReduce and SQL

Data Application Lab

Motivation

- Spot the difference?



Intuitive

- Make unstructured data looks like tables regardless how it really lays out
- SQL-like query can operate directly against these tables
- Less development time
- Easy for adhoc analysis
- Place for multi-user, multi-session
- User-friendly

Data Application Lab

What Is Hive

- **Data warehouse** infrastructure built on top of **Hadoop** for providing data summarization, query and analysis

- What is Data Warehouse?
  - Is a database specific for analysis and reporting purposes

- Designed for OLAP
- Provides SQL type language for querying called **Hive QL**
- Familiar, fast, scalable and extensible

How Hive Works

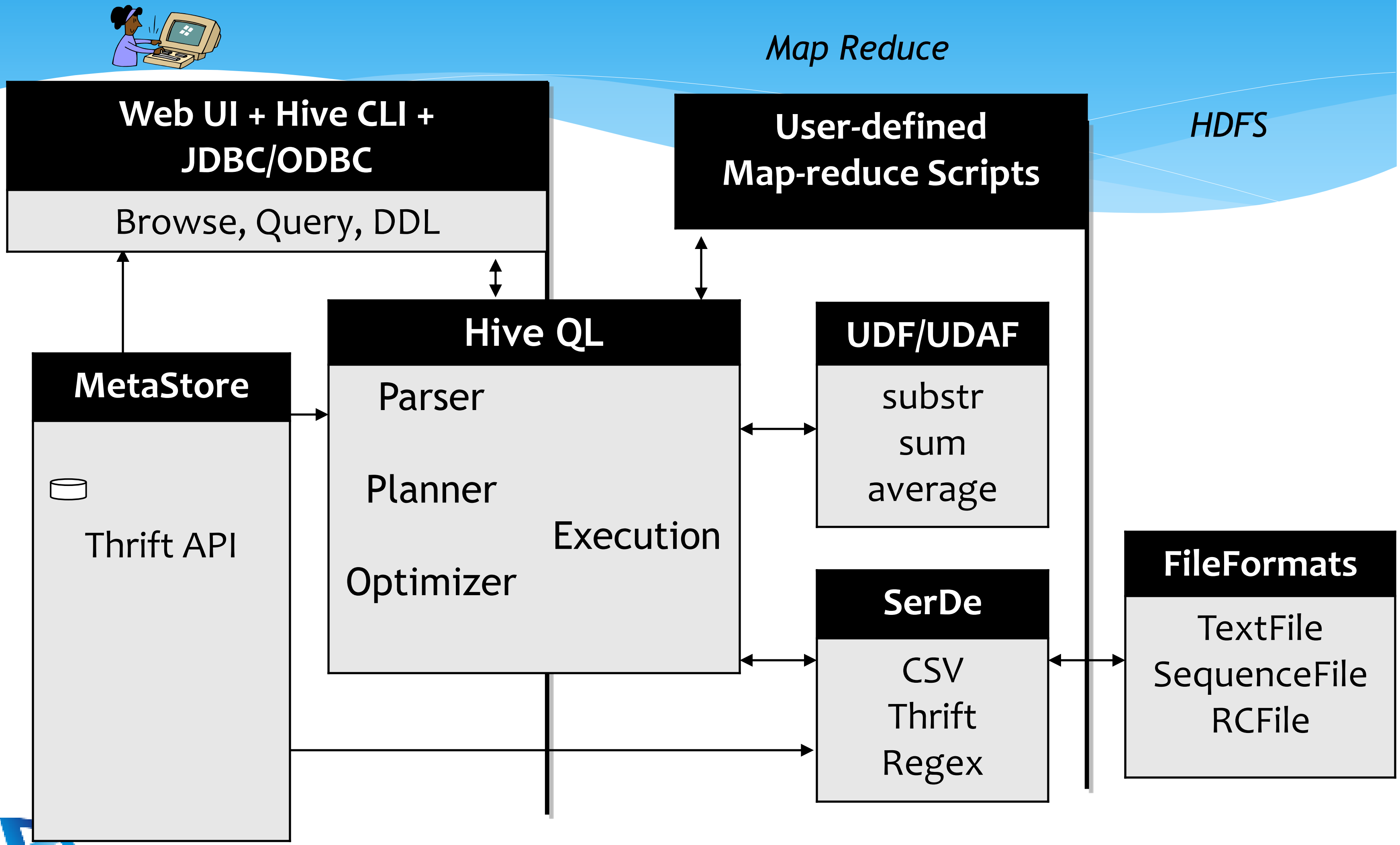
- Hive built on top of Hadoop
  - Think HDFS and MapReduce
- Hive stored data in the HDFS
- Hive compile SQL queries into MapReduce jobs and run the jobs in the Hadoop cluster

Data Application Lab

What Hive Is NOT

- Not work with small data set (high latency)
- Not designed for online transaction processing
- Not offer real-time queries
- Not work as row level query

Hive Architecture



Data Application Lab

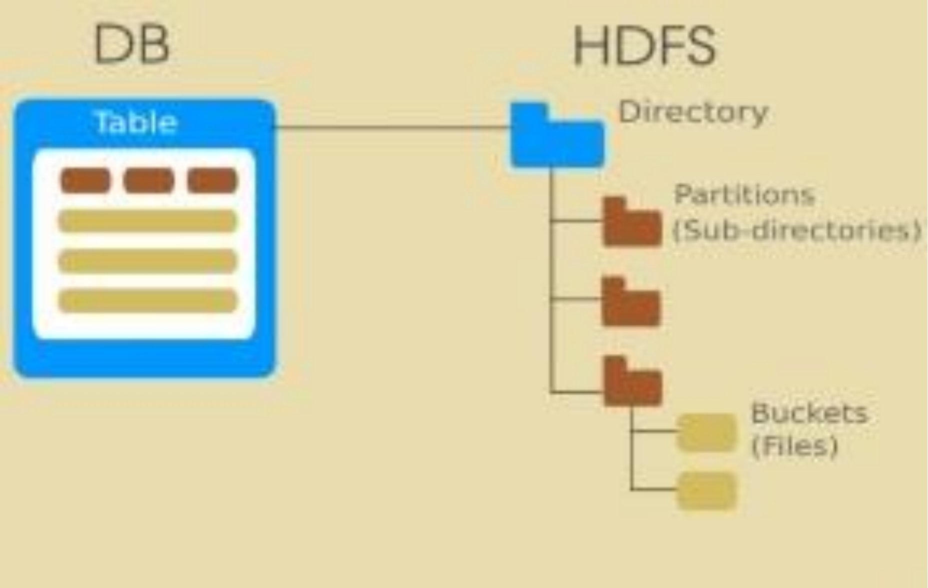
Learn Hive

- Hive Data Model
- Query Language

Data Application Lab

Hive Data Model

- Hive structure data into a well defined database concept i.e. tables, columns, and rows, partitions, buckets etc



Data Application Lab

Hive Data Model

- Tables
  - Basic type columns(int, float, boolean)
  - Complex types: array/map/struct/…

```
CREATE TABLE employee( id INT,  name STRING);
```

- Partitions
  - i.e. range partition tables by date

```
CREATE TABLE sales( id INT, items  STRING) PARITIONED BY (ds STRING);
```

- Buckets
  - Useful for sampling

Data Application Lab

Metadata

- Database
  - Namespace containing a set of tables
- Table definitions
  - Contains list of columns and their types and SerDe Info
  - Schema info, physical location on HDFS
- Partition
  - Each partition can have its own columns and storage info
- Statistics
  - Info about the databases

Data Application Lab

Hive Physical Layout

- Warehouse directory in HDFS
- Table row data is stored in warehouse sub-directories
- Partition creates sub-directories within table directories

Creation of a Table on Hive

- hive > CREATE TABLE new_students(ID INT, studentName STRING);
- hive > CREATE TABLE students(ID INT, studentName STRING) ROW FORMAT DELIMITED FIELDS TERMINATED BY ', ';

Data Application Lab

Load data into a Hive Table

- Hive does not do any transformation while loading data into tables
- Load operations are currently pure copy/move operations that move data files into locations corresponding to hive tables
- Example
  - *LOAD DATA LOCAL INPATH '/demo/students_tb/students/students.txt' OVERWRITE INTO TABLE students;*

Data Application Lab

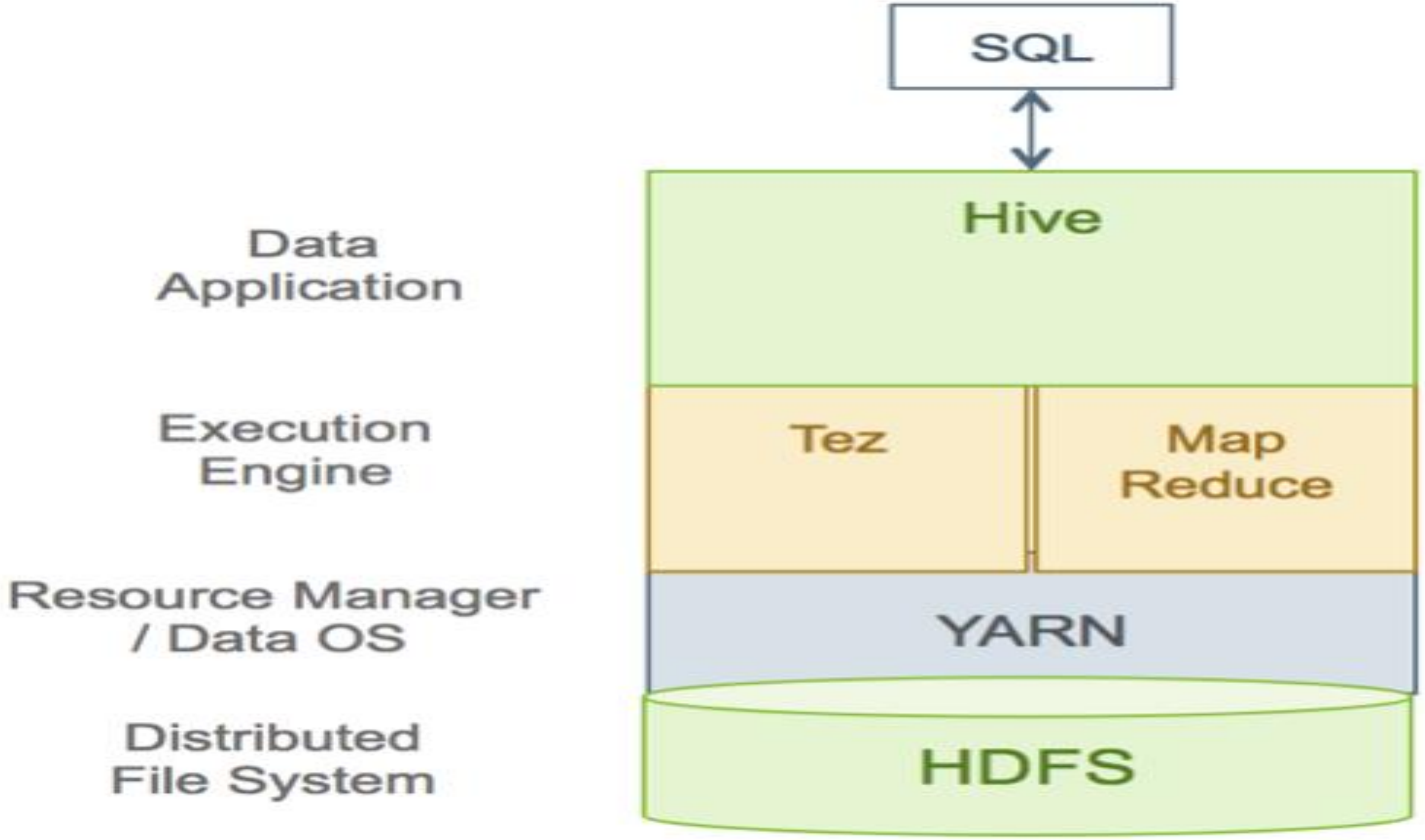
Hive Execution

- CLI: Hive CLI & Beeline
- Web UI: Hue, Ambari
- JDBC/ODBC

- Run one query
  - *hive -e 'SELECT DISTINCT username FROM temp.TwitterExampletextexample LIMIT 10'*
- Run a hive query file
  - *hive -f /tmp/demo_hive.sql*

Data Application Lab

SQL Query Execution Process

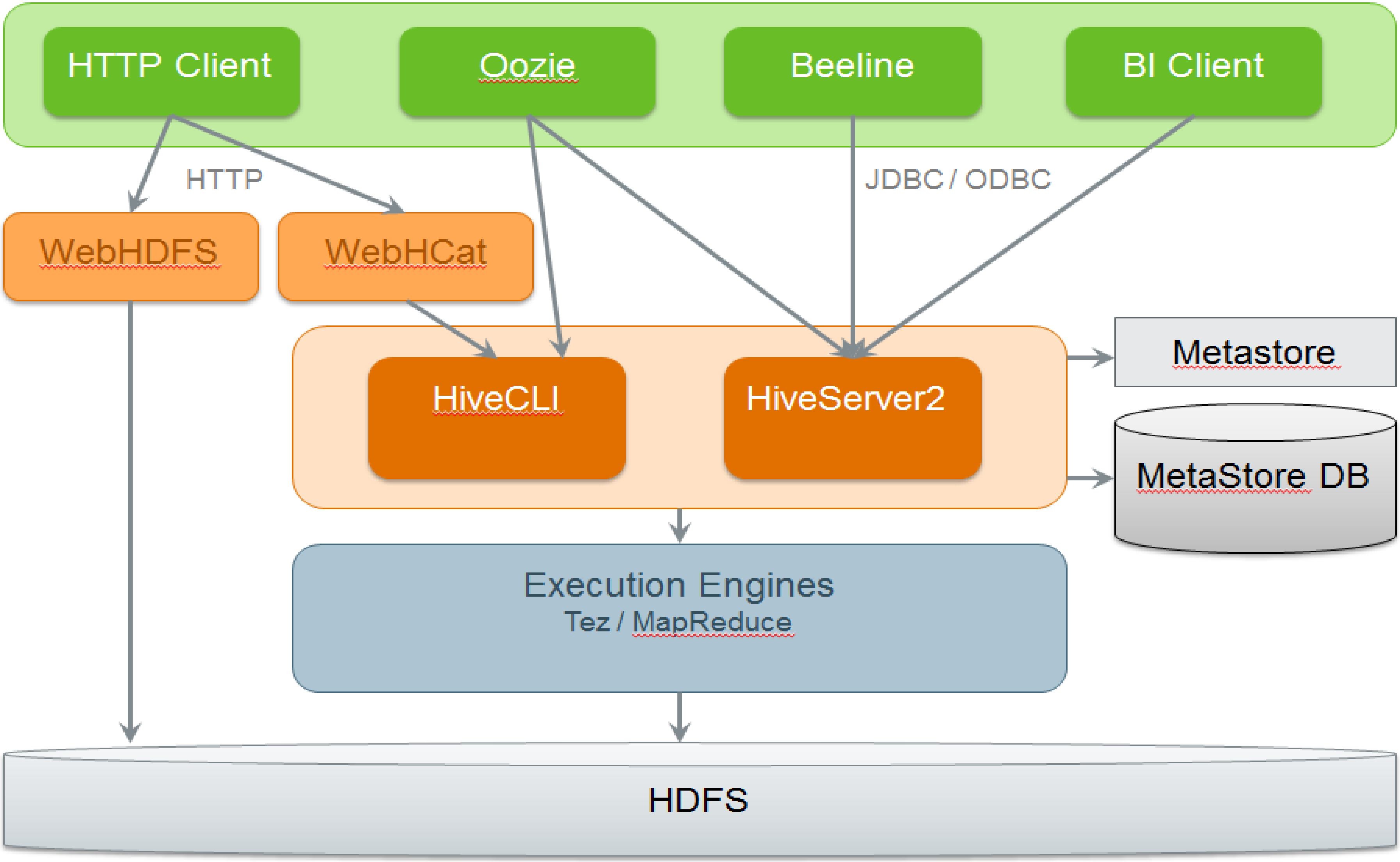


Data Application Lab

Data Type (primitive)

- TINYINT
- SMALLINT
- INT
- BIGINT
- BOOLEAN
- FLOAT
- DOUBLE
- BIGDECIMAL
- STRING
- BINARY
- TIMESTAMP

Hive Stack


HTTP Client
Oozie
Beeline
BI Client
HTTP
JDBC / ODBC
WebHDFS
WebHCat
HiveCLI
HiveServer2
Metastore
MetaStore DB
Execution Engines
Tez / MapReduce
HDFS

Hive Database Features

- All about files
- Schema on read
- Fast when load data into DB
- Touch data only when run query
- Don’t support delete/update

Data Application Lab

Data Units

- Database
- Table
- Partition – are done on columns
  - *CREATE TABLE students_part(ID INT, Name STRING) PARTITIONED BY (dept STRING)*
  - Above command create a sub-directory for each value of the partition column
    - */user/hive/warehouse/stutents_part/dept=cs/*
  - Queries with partition columns in WHERE clause will scan through only a subset of data

Data Application Lab

Database

- Create
- Use
- Drop

```
CREATE (DATABASE|SCHEMA) [IF NOT EXISTS] database_name
  [COMMENT database_comment]
  [LOCATION hdfs_path]
  [WITH DBPROPERTIES (property_name=property_value, ...)];
```

Data Application Lab

Table

- External vs internal
- Create
- Drop

```
CREATE [TEMPORARY] [EXTERNAL] TABLE [IF NOT EXISTS] [db_name.]table_name    -- (Note: TEMPORARY available in
Hive 0.14.0 and later)
 [(col_name data_type [COMMENT col_comment], ...)]
 [COMMENT table_comment]
 [PARTITIONED BY (col_name data_type [COMMENT col_comment], ...)]
 [CLUSTERED BY (col_name, col_name, ...) [SORTED BY (col_name [ASC|DESC], ...)] INTO num_buckets BUCKETS]
 [SKEWED BY (col_name, col_name, ...)                -- (Note: Available in Hive 0.10.0 and later)]
   ON ((col_value, col_value, ...), (col_value, col_value, ...), ...)
   [STORED AS DIRECTORIES]
 [
  [ROW FORMAT row_format]
  [STORED AS file_format]
   | STORED BY 'storage.handler.class.name' [WITH SERDEPROPERTIES (...)]  -- (Note: Available in Hive 0.6.0 and later)
 ]
 [LOCATION hdfs_path]
 [TBLPROPERTIES (property_name=property_value, ...)]   -- (Note: Available in Hive 0.6.0 and later)
 [AS select_statement];   -- (Note: Available in Hive 0.5.0 and later; not supported for external tables)
```

Basic Queries

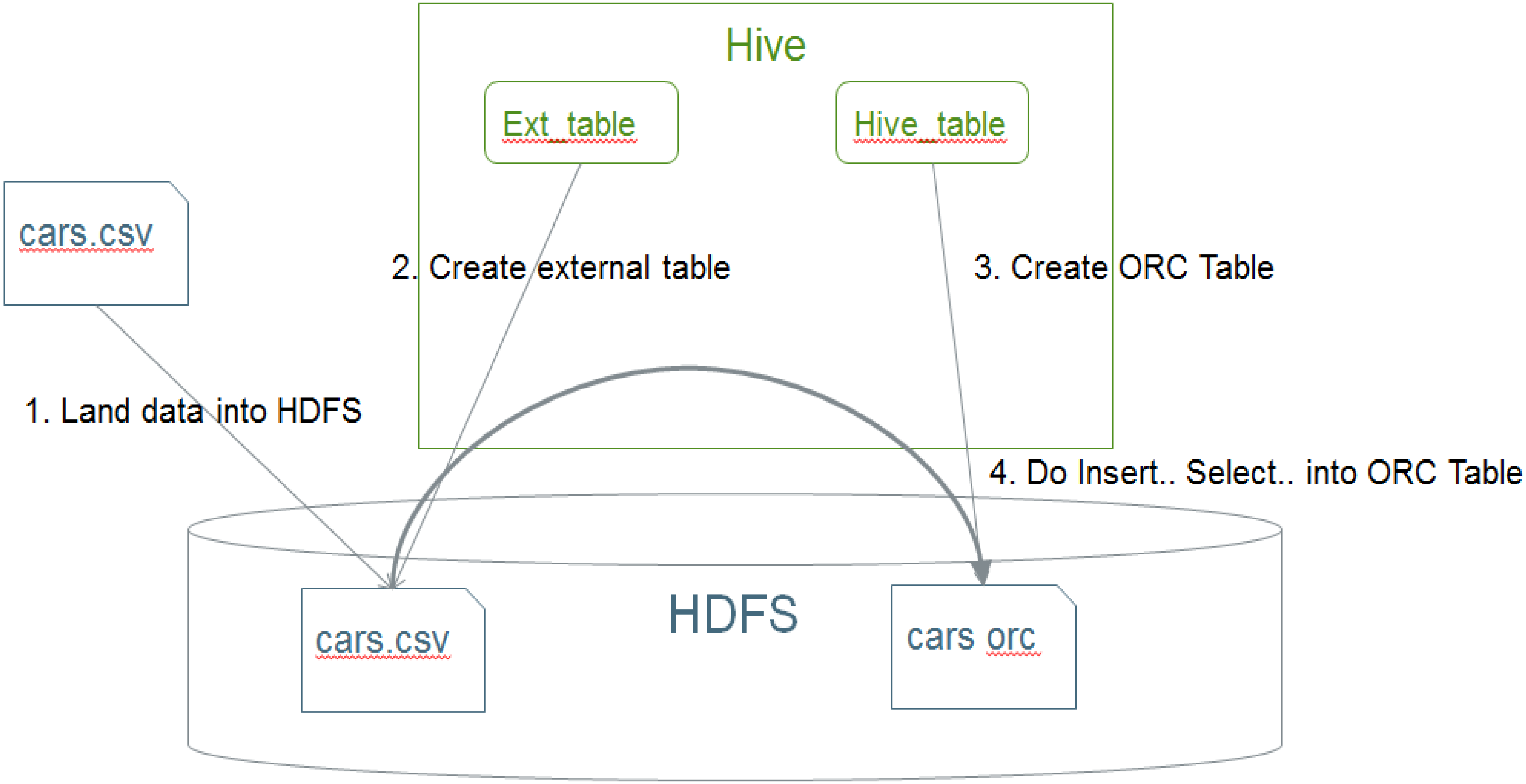
- Select
- Count
- Join
- Group

LOAD DATA INPATH '*hdfs_file_or_directory_path*' [OVERWRITE] INTO TABLE *tablename* [PARTITION (*partcol1=val1*, *partcol2=val2* ...)]

Best Practice to Use Hive

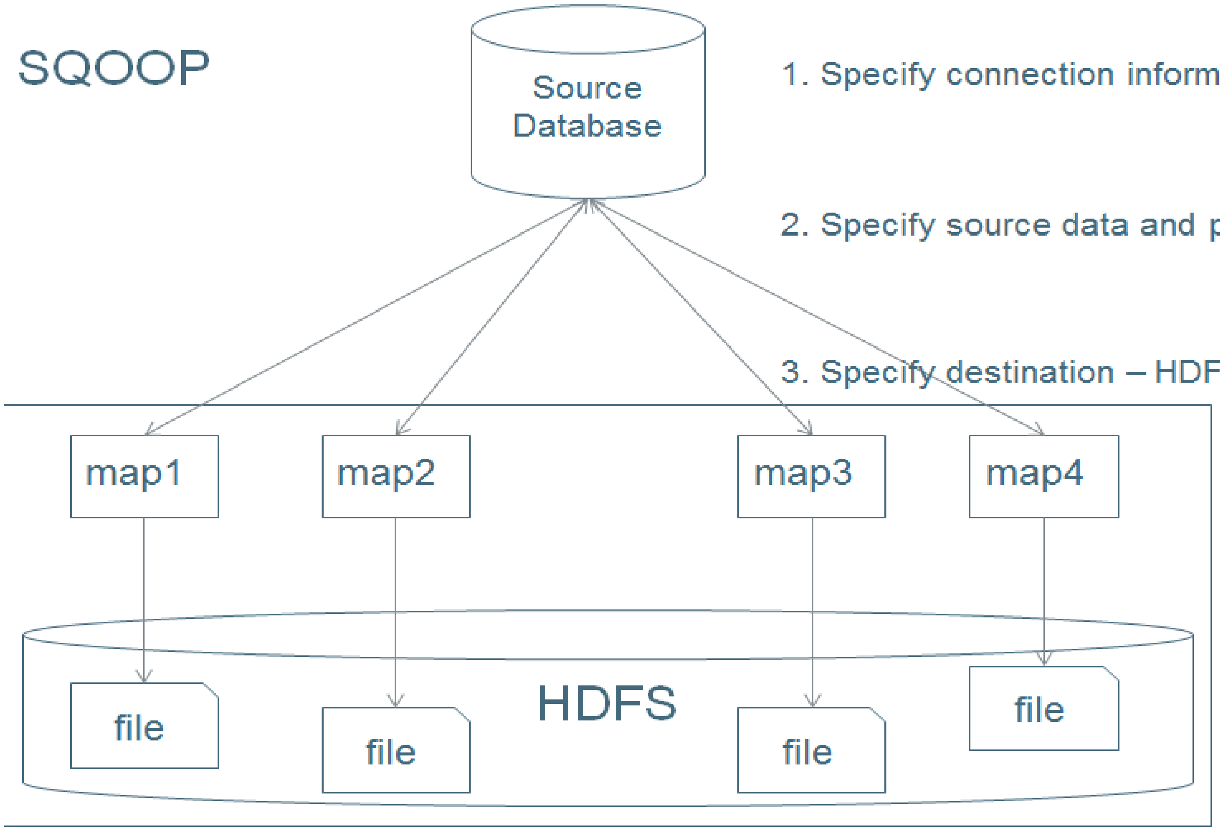
- Getting Data into Hive
- Storing Data In Hive
  - Correct Storage is the key to performance
- Execution Engine
- Optimize Query

Hive Ingestion: Using External Table



Data Application Lab

Hive Ingestion: Sqoop



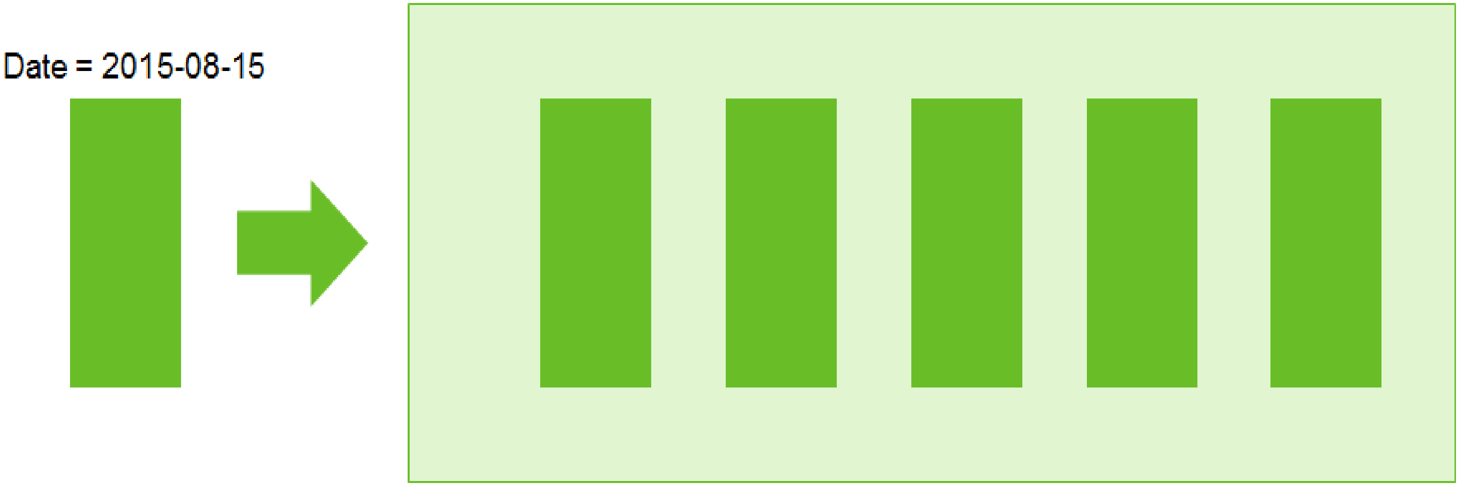
1. Specify connection information

2. Specify source data and parallelism (=4 here)

3. Specify destination – HDFS or Hive (prefer HDFS)

Store Data Using Partitions

1. Primary
   - Atomicity of Append
   - Reducing search space for Query
2. Secondary
   - Reduce space for compactions
   - Reduce space for updates (partition replacement)



Note: Schema evolution is supported on partitions without changing old data, However you cannot modify old partitions if the schema changes

Data Application Lab

High Performance Hive

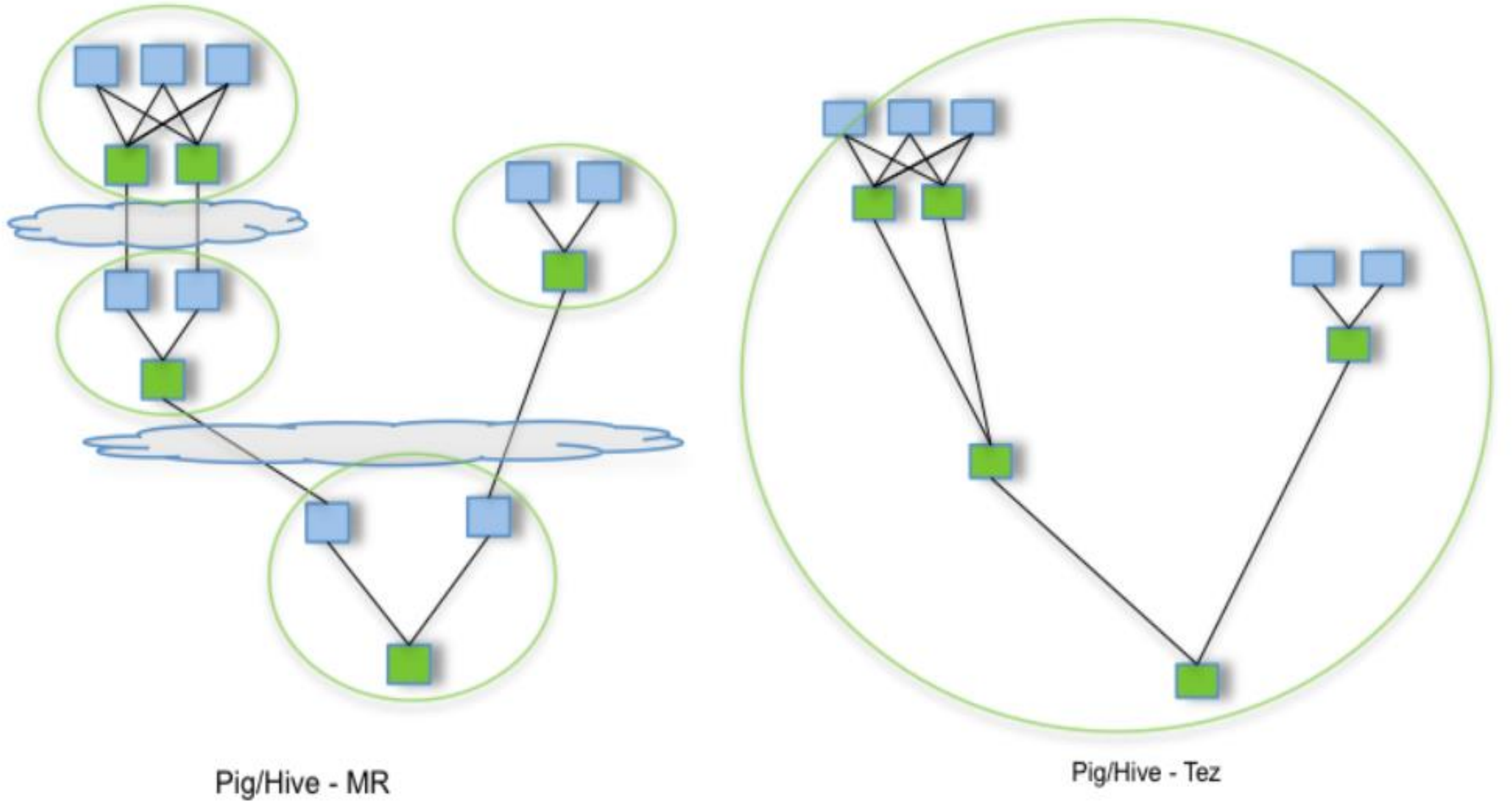
- Use the ORC/parquet File Format
- Use the Tez Query Execution Engine
- Use proper Compression techniques
- Better Workload Management by Using Queues

Data Application Lab

Advanced Columnar format

- Use the RC/ORC/parquet File Format
- Record Columnar(RC) format determines how to store relational tables on distributed computer clusters.
- Optimized Row Columnar (ORC) File format is used as it further compresses data files.
- Parquet – latest standard of columnar storage format for Big data storage

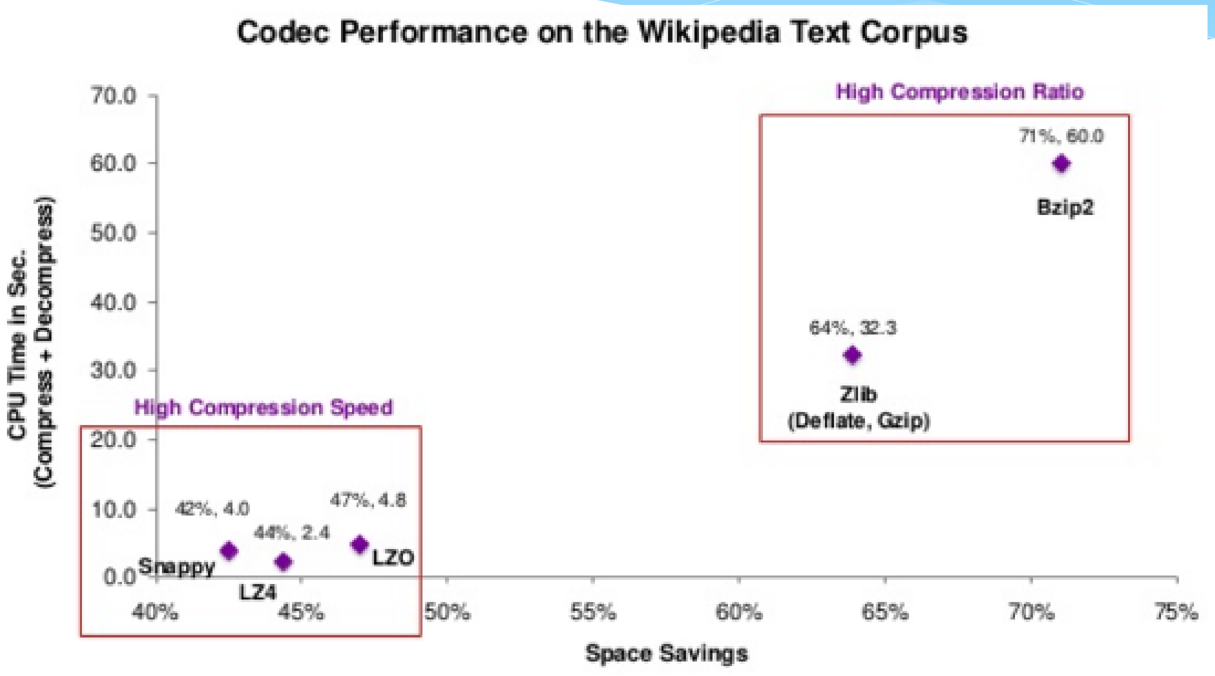
Use Tez Execution Engine



Data Application Lab

Compression

- Space or



- Splittable?

Data Application Lab

HiveQL and Spark

- What is Spark
- Dataframe (equivalent to a table)
- Use HiveQL in Spark

```
import org.apache.spark.sql.hive.HiveContext
val hiveContext = new HiveContext(sc)
val results = hiveContext.sql(”… ")
```

HiveQL and Spark, Con’t

```
//  prep
import org.apache.spark.sql.hive.HiveContext
val sc = new SparkContext(new SparkConf().setAppName(this.getClass.getName))
val hiveContext = new HiveContext(sc)

//  Create a table
hiveContext.sql("create table ratings ( ”
"user_id int, ”
"movie_id int, ”
"rating int, ”
"ts bigint) ”
"row format delimited ”
"fields terminated by ’*’ ”
"lines terminated by ’\n’ ”
"stored as textfile")
```

HiveQL and Spark, Con’t

```
//  load data into a table
hiveContext.sql("load data inpath
’/data/movielens_1m_simple/ratings/ratings.dat’ ”
"into table ratings")

//query on this table
Val genres = hiveContext.sql("select explode(genres) as genre ”
"from movies")
genres.take(10)
```

Use cases

- Log processing
  - Daily Report
  - User Activity Measurement
- Data/Text mining
  - Machine learning (Training Data)
- Business intelligence
  - Advertising Delivery
  - Spam Detection
- Predictive Modeling, Hypothesis Testing

Most Asked Interview Questions

- Types of Hive tables? How are they different?
- Is Hive suitable for OLTP
- What is metastore in Hive
- Why we need Hive
- How do we get HDFS location for a table
- How do you check a partition
- What is the significance of ‘IF EXISTS’/ “IF NOT EXISTS”
- When you point a partition of a hive table to a new directory, what happens to the data?
- While loading data into a hive table using the LOAD DATA clause, how do you specify it is a hdfs file and not a local file ?

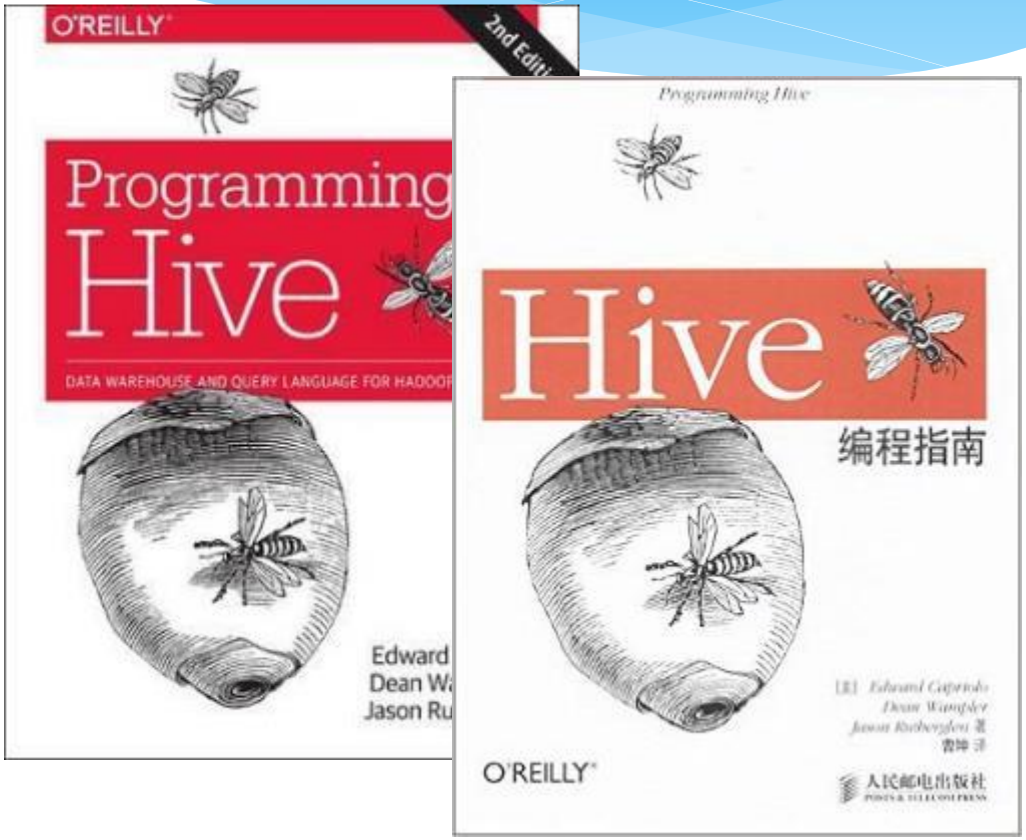
Future Roadmap

- Hive 2.1.1 – released Dec, 2016
- Hive on Spark

Data Application Lab

Recommended reading

- Book
- hive.apache.org



Data Application Lab

Home Project

- Reproduce demo projects
- TA Project