cloudera
Ask Bigger Questions

Apache Hive – Best Practices

Mark Grover
Software Engineer, Cloudera Inc.
www.github.com/markgrover/bdtc-hive
@mark_grover

Logistics

- Hadoop and Hive installed
  - Method 1: Demo VM at https://bitly.com/PDfLKp
  - Method 2: Install Hadoop + Hive (>= 0.7)
- Dataset downloaded
- All instructions at github.com/markgrover/bdtc-hive

Agenda

- Interactive
- Frequent breaks and conversations!
- Learn a little about Hive

My request to you

- Follow along
- Ask questions, any questions

Hive

- What is Hive?
- SQL-like queries for processing Hadoop data

Hadoop

- What is Hadoop?
- Run some example MR jobs

The problem with MapReduce

- Catered towards developers
- No higher level abstraction

Hive

- Why use Hive?
- Scalable
- Select, where clause, group by, order by, joins
- Pluggable User Defined Functions - UDFs (e.g from_unixtime)
- Pluggable User Defined Aggregate Functions - UDAFs (e.g. count, avg)
- Pluggable User Defined Table Generating Functions - UDTFs (e.g. explode)

Hive...2

- Pluggable custom Input/Output format
- Pluggable Serialization Deserialization libraries (SerDes)
- Pluggable custom map/reduce scripts

What Hive does NOT support

- OLTP workloads - low latency
- Correlated subqueries

Let's get started

- Load the data on HDFS
- Create a Hive table
- Run some queries
- Connecting to Hive
  - Shell
  - JDBC
  - ODBC
  - Thrift client

Inside Hive

- Configuring Hive
- Hive metastore
- Partitioning
- Dynamic partitioning
- Bucketing
- Joins
- UDFs/UDAFs
- Custom map/reduce scripts
- Hive HBase Integration

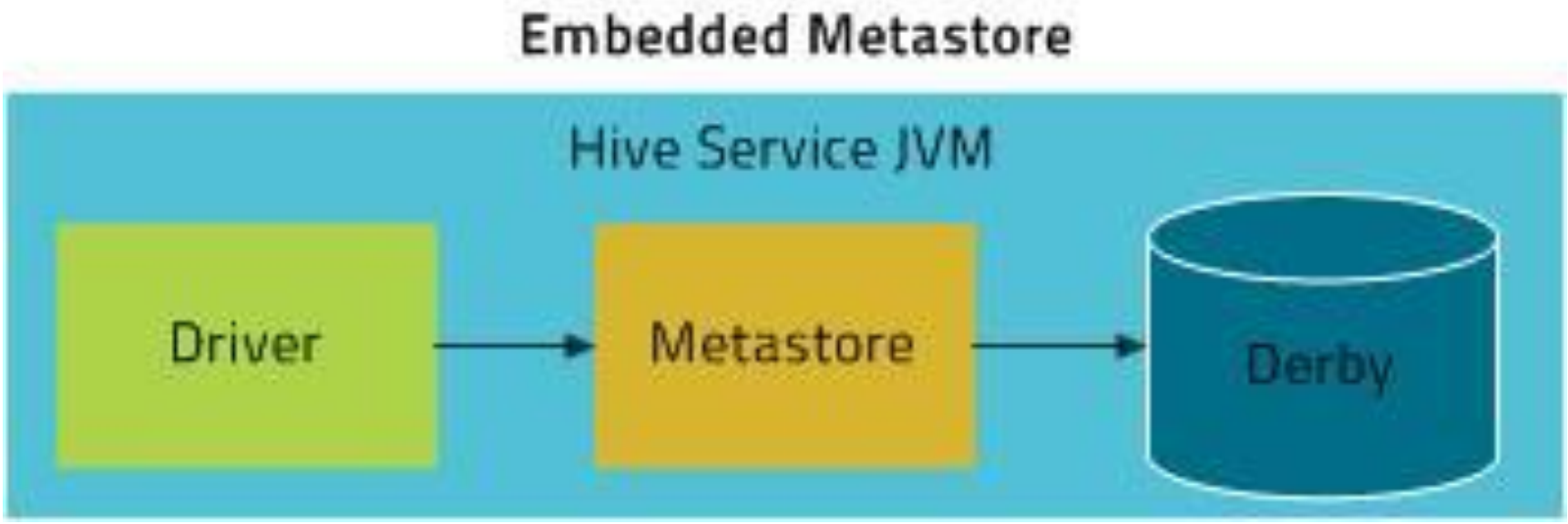
Configuring Hive

- hive-default.xml
- hive-site.xml (/etc/hive/conf/hive-site.xml)

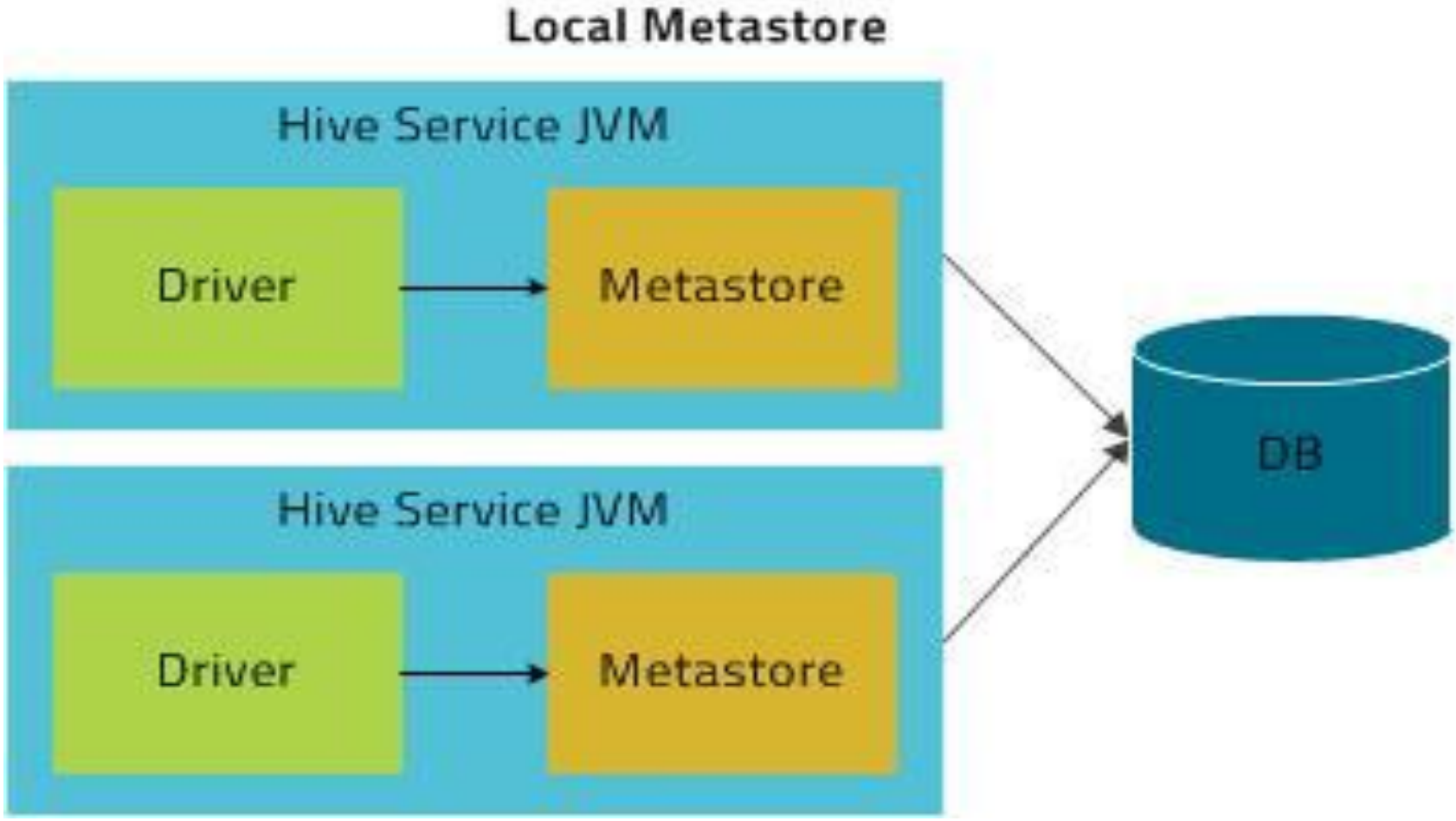
Hive metastore

- What is metastore?
- Vocabulary
  - Metastore
  - Metastore server (or datastore)
- Metastore service runs in different modes
  - Embedded
  - Local
  - Remote
- Default is embedded derby
  - Not recommended for anything non-trivial

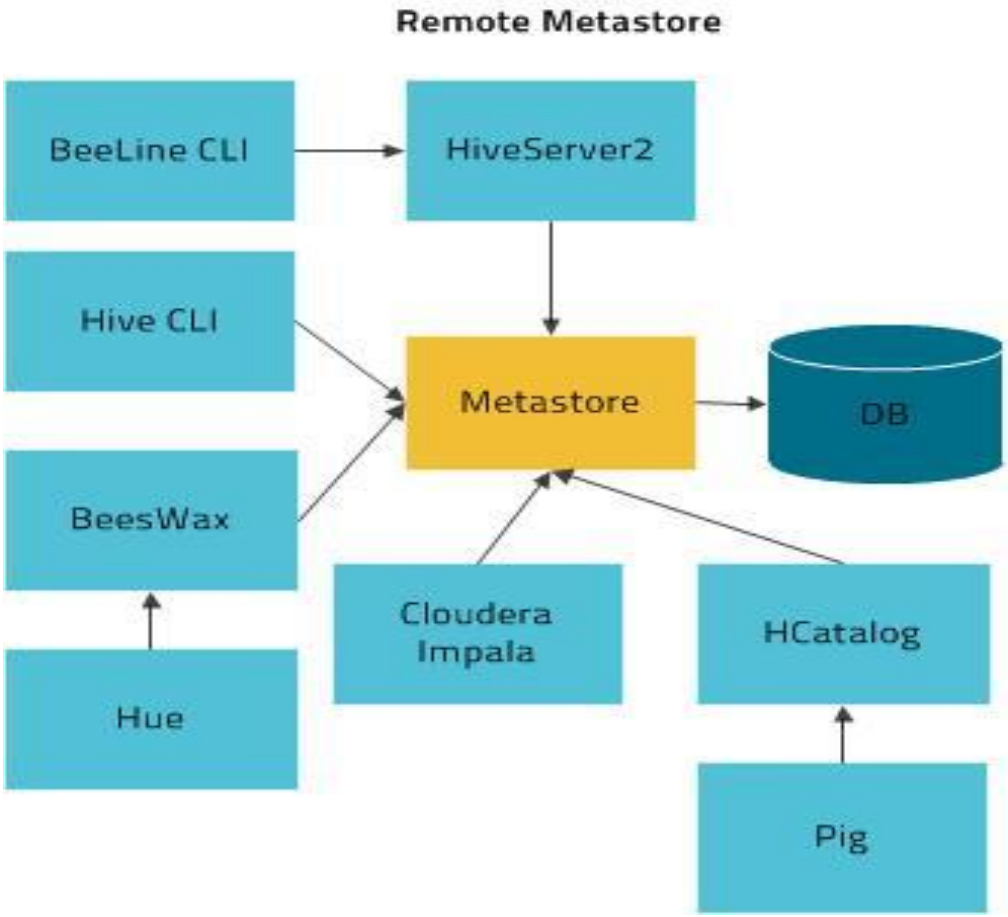
Embedded Metastore Server



Local Metastore Server



Remote Metastore Server



Configuring Hive metastore

- Default - embedded derby

```
<property>
 <name>javax.jdo.option.ConnectionURL</name>
<value>jdbc:derby:;databaseName=metastore_db;create=true</value>
 <description>JDBC connect string for a JDBC metastore</description>
</property>
```

Partitions in Hive

- Sub-directories under table directories
- Partition pruning

Partitions in Hive

- Create a partitioned table
- Run a query on the partitioned table and compare the time with non-partitioned table

Partitioning best practices

- Why/When to partition?
  - WHERE, GROUP BY, JOINS benefit
- Which column to partition by?
- How many partitions?
  - Don’t overload the metastore (<10,000 partitions)
  - Don’t run into small files problem

Dynamic partitioning

- Populate a partitioned table using Dynamic Partitioning

Joins in Hive

- Regular (aka common) joins
  - Reduce side
- Map joins
- Bucketed Map joins
- Sorted Bucketed Merge (SMB) joins

Joins: best practices

- When to use which join?
- Regular join

Bucketing in Hive

- Hash based bucketing within a partition
- For faster joins
- For sampling
- Sorted buckets

Sampling in Hive

- You only want a sample of the data
- Sampled on a particular column

Bucketing in Hive

- Create a bucketed table
- Perform a sampling query on this table

Bucketing best practices

- When to use bucketing?
- Which columns to bucket on?
- How many buckets to use?

UDFs/UDAFs

- What is a UDF?
- Custom Functionality
  - written in Java

Custom MR scripts

- Plug in custom MapReduce scripts in Hive queries

Custom plugin best practices

- When to use UDF and when to use custom MR scripts?
- Comparison between UDFs/UDAFs and custom MR scripts

Hive-HBase Integration

- Read from and write to HBase tables using Hive queries

Contact info

@mark_grover

github.com/markgrover

linkedin.com/in/grovermark

mgrover@cloudera.com

Office hours:

Cloudera booth from 3 - 4pm tomorrow!