Introduction to Apache Hive (and Hcatalog)

Mark Grover

[github.com/markgrover/nyc-hug-hive]()

Me!

- Contributor to Apache Hive
- Section Author of O'Reilly's Programming Hive book
- Software Developer at Cloudera
- @mark_grover
- mgrover@cloudera.com
- https://github.com/markgrover/nyc-hug-hive

Agenda

- What is Hive?
- Why use Hive?
- Hive features
- Hive architecture
- HCatalog
- Demo!

Preamble

- This is a remote talk
- Feel free to ask questions any time!

Agenda

- What is Hive?
- Why use Hive?
- Hive features
- Hive architecture
- HCatalog
- Demo!

Hive

- Data warehouse system for Hadoop
- Enables Extract/Transform/Load (ETL)
- Associate structure with a variety of data formats
  - Persisted in Hive metastore
- Access to files in HDFS, HBase, etc.
- Query execution in MapReduce

Agenda

- What is Hive?
- Why use Hive?
- Hive features
- Hive architecture
- HCatalog
- Demo!

Why use Hive?

- MapReduce is catered towards developers
- Run SQL-like queries that get compiled and run as MapReduce jobs
- Data in Hadoop even through generally unstructured has some vague structure associated with it
- Benefits of MapReduce + Hadoop
  - Fault tolerant
  - Robust
  - Scalable

Agenda

- What is Hive?
- Why use Hive?
- Hive features
- Hive architecture
- HCatalog
- Demo!

Hive features

- Create table, create view, create index - DDL
- 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 features

- 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
- Not super performant with small amounts of data
  - How much data do you need to call it “Big Data”?

Other Hive features

- Partitioning
- Sampling
- Bucketing
- Various kinds of optimized joins
- Integration with HBase and other storage handlers
- Views – Unmaterialized
- Complex data types – arrays, structs, maps

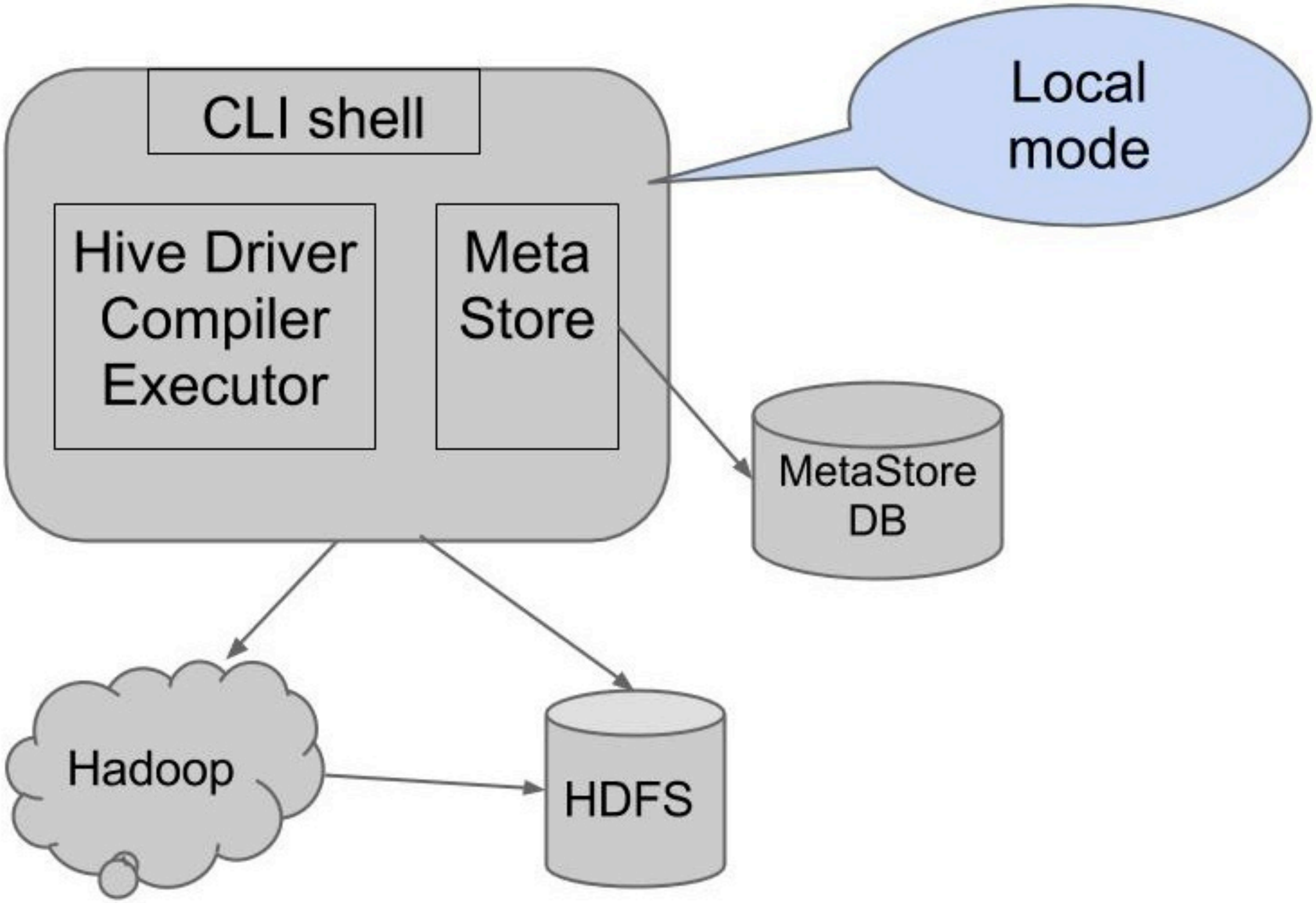
Connecting to Hive

- Hive Shell
- JDBC driver
- ODBC driver
- Thrift client

Agenda

- What is Hive?
- Why use Hive?
- Hive features
- Hive architecture
- HCatalog
- Demo!

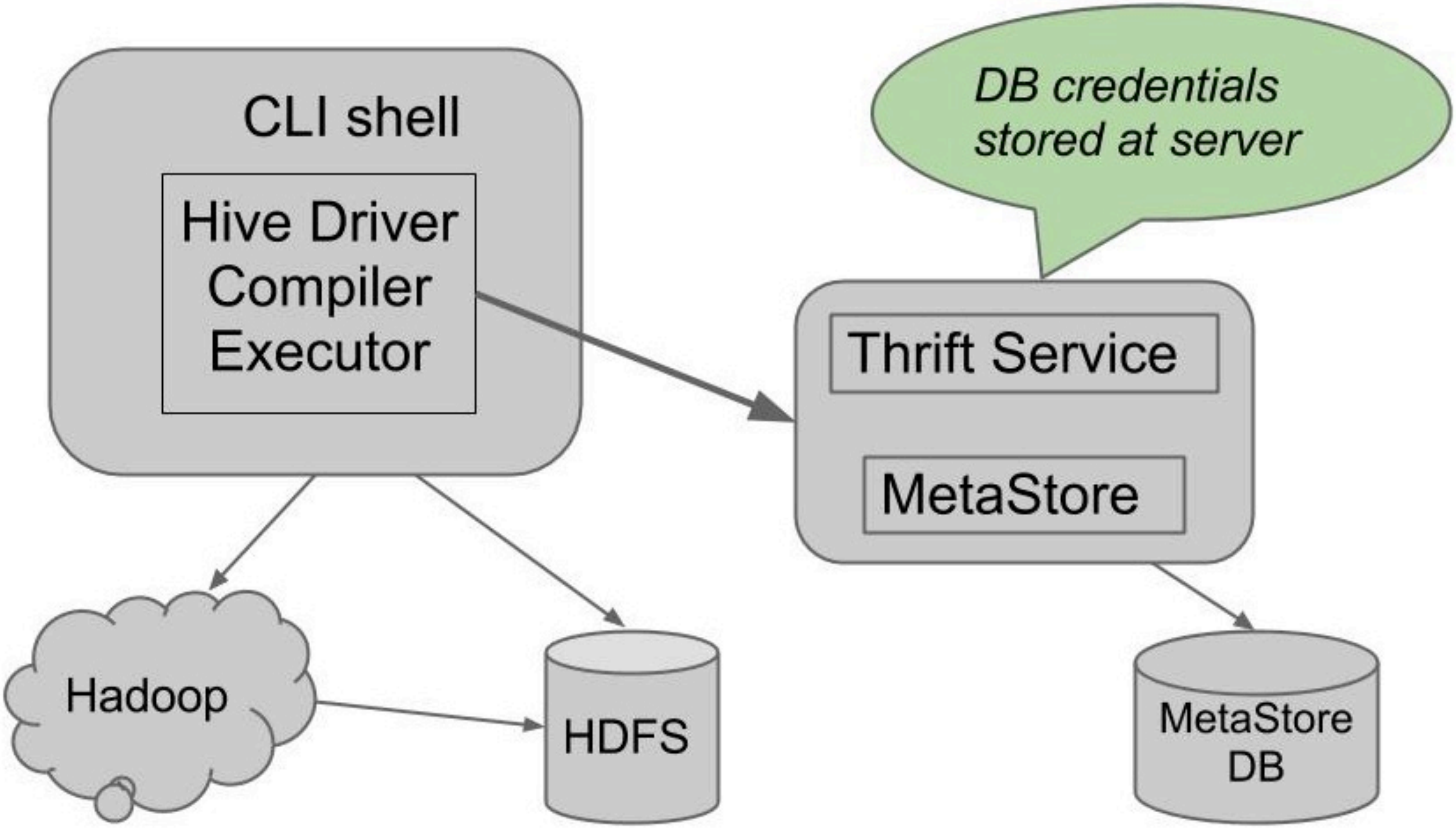
Hive architecture



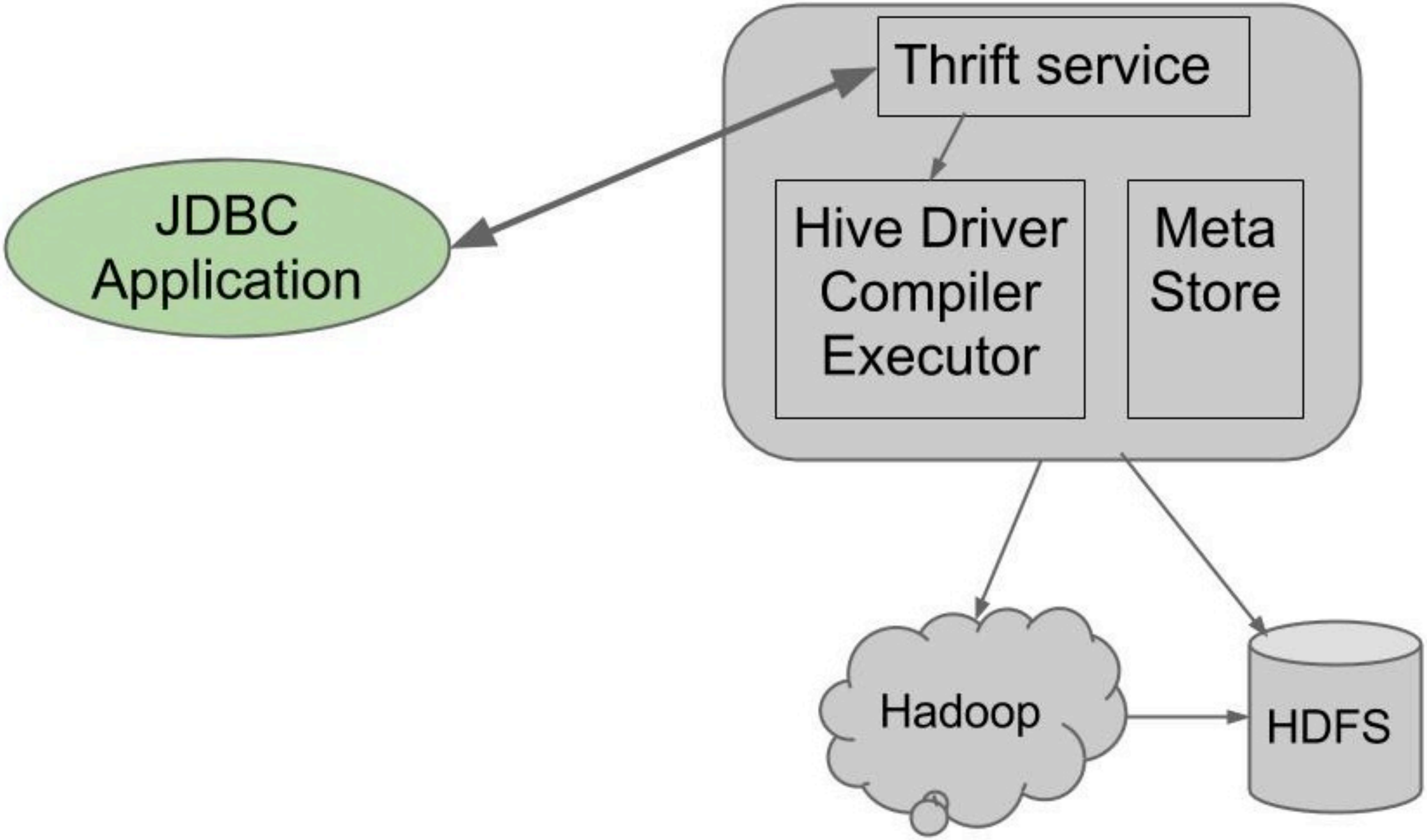
Hive metastore

- Persists schema
- Default Embedded Derby
  - Not recommend for anything but a quick Proof of Concept
- 3 different modes of operation:
  - Embedded Derby (default)
  - Local
  - Remote

Hive Remote Mode



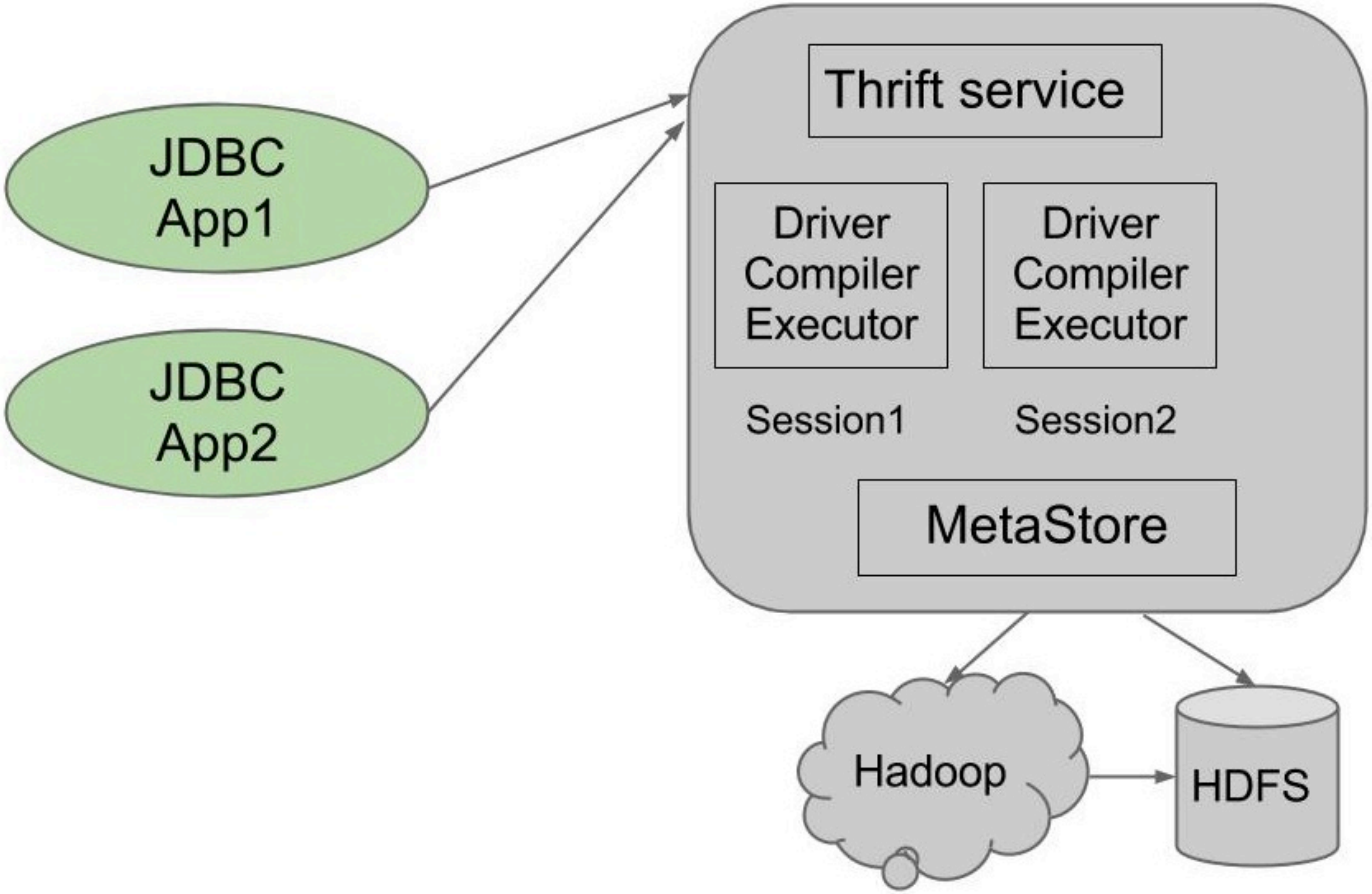
Hive server


Thrift service
Hive Driver
Compiler
Executor
Meta
Store
JDBC
Application
Hadoop
HDFS

Problems with Hive Server

- No sessions/concurrency
- Essentially need 1 server per client
- Security
- Auding/Logging

Hive server 2



Hive architecture

- Compiler
  - Parser
  - Type checking
  - Semantic Analyzer
  - Plan Generation
  - Task Generation

Hive architecture

- Execution Engine
  - Plan
  - Operators
  - SerDes
  - UDFs/UDAFs/UDTFs
- Metastore
  - Stores schema of data
  - HCatalog

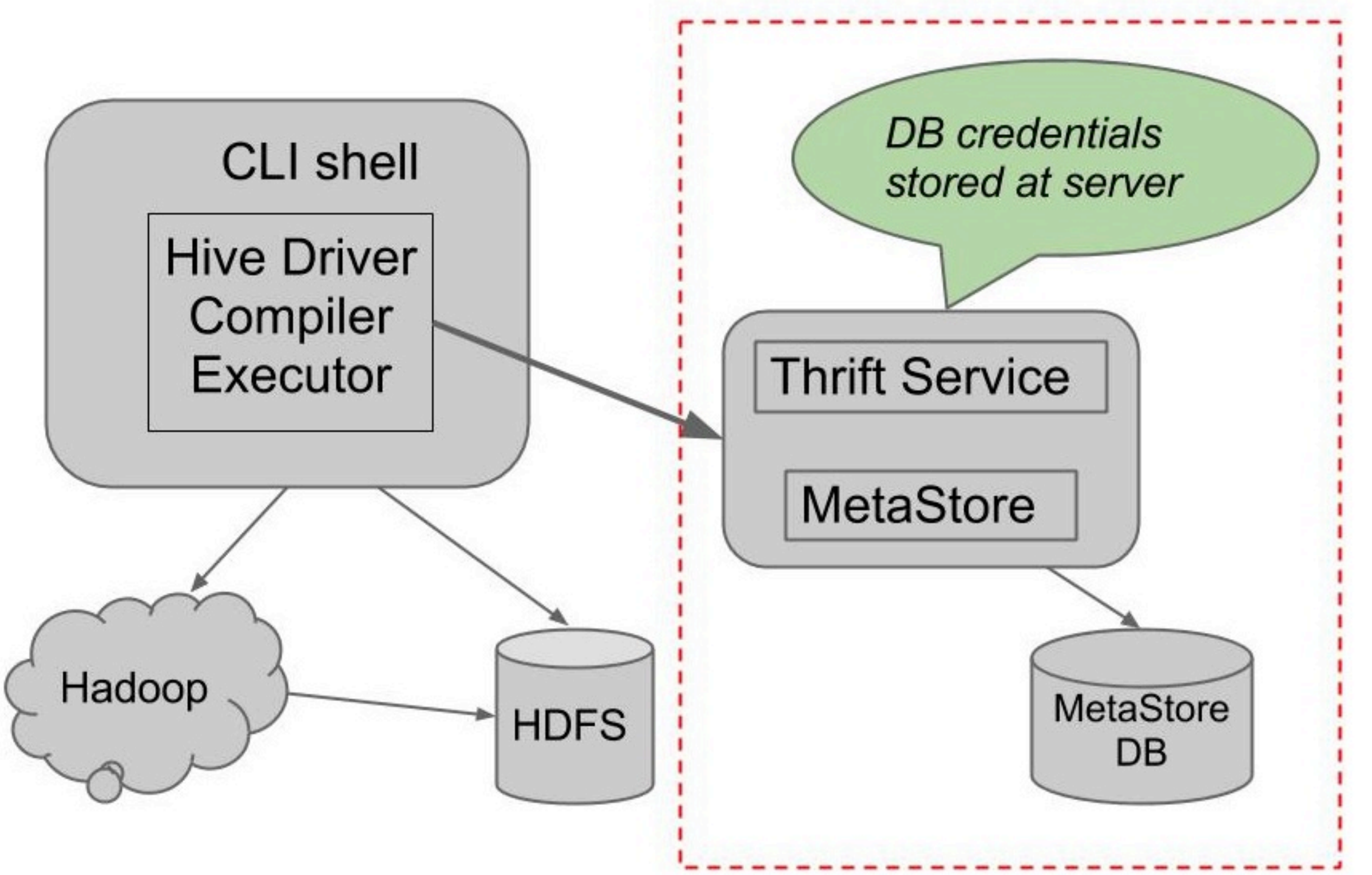
Architecture Summary

- Use remote metastore service for sharing the metastore with HCatalog and other tools
- Use Hive Server2 for concurrent queries

Agenda

- What is Hive?
- Why use Hive?
- Hive features
- Hive architecture
- HCatalog
- Demo!

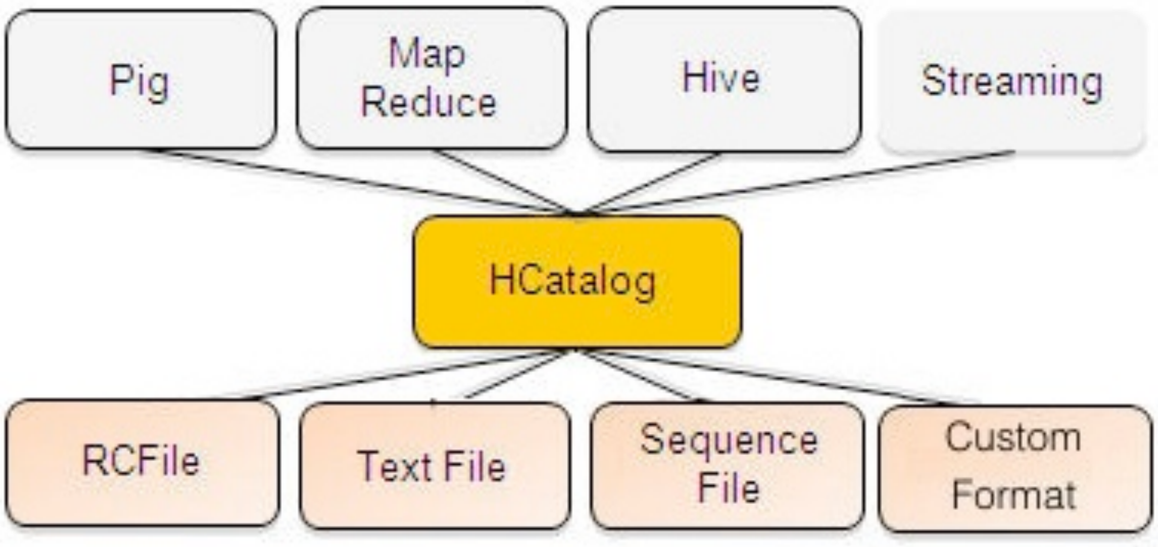
Hive Metastore Remote Mode



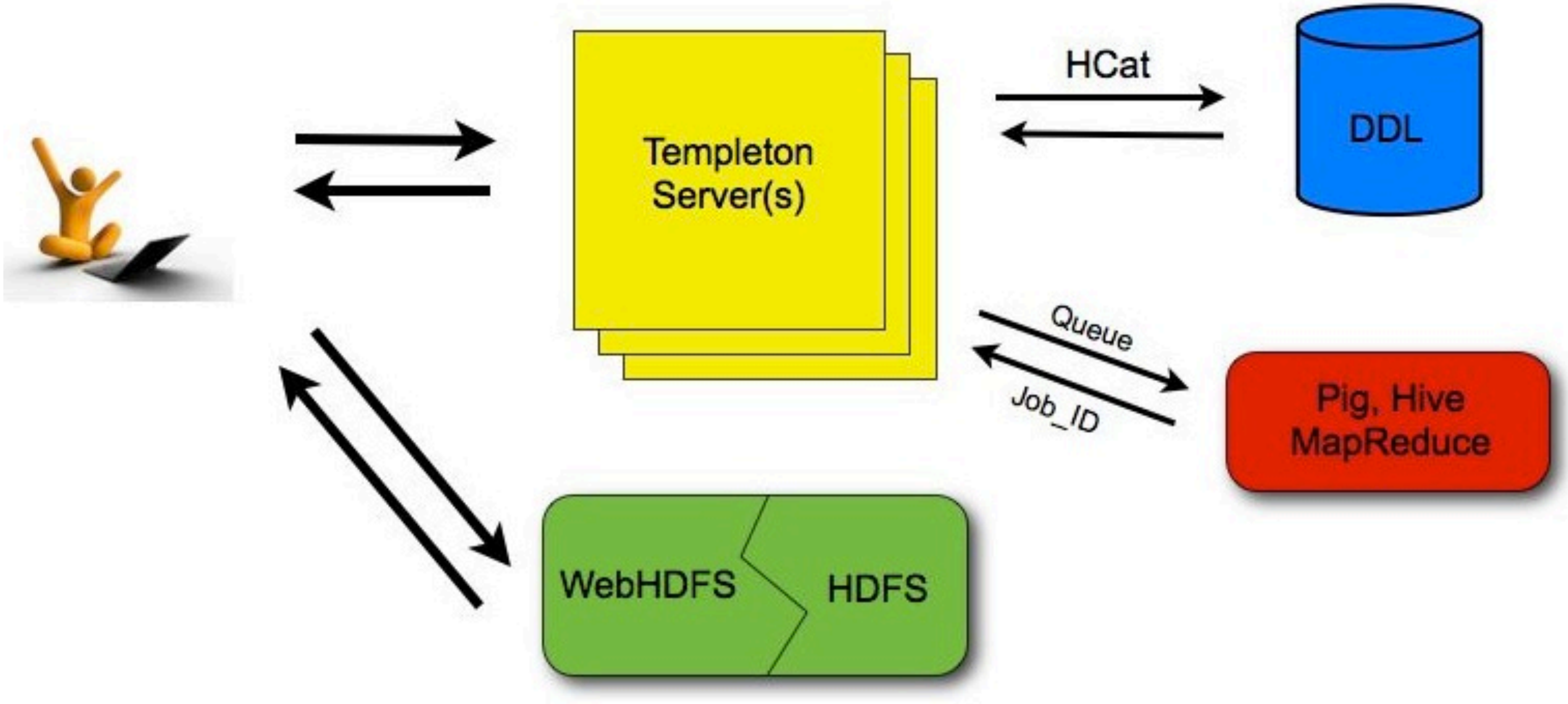
HCatalog

- Table and storage management service
- Metastore contains information of interest to other tools (Pig, MapReduce jobs)
- Expose that information as REST interface
- WebHCat: Web Server for engaging with the Hive metastore

HCatalog


Pig
Map Reduce
Hive
Streaming
HCatalog
RCFile
Text File
Sequence File
Custom Format

WebHCat


Templeton
Server(s)
HCat
DDL
Queue
Job_ID
Pig, Hive
MapReduce
WebHDFS
HDFS

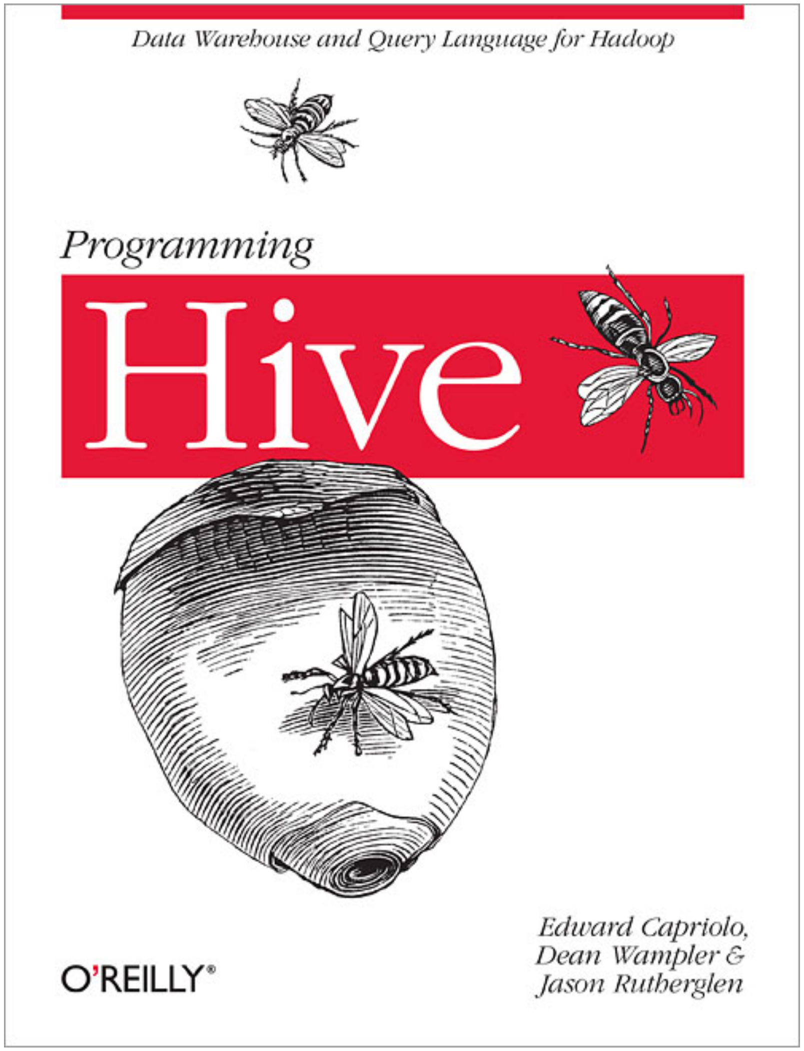
Agenda

- What is Hive?
- Why use Hive?
- Hive features
- Hive architecture
- HCatalog
- Demo!

Applications of Hive

- Web Analytics
- Retail
- Healthcare
- Spam detection
- Data Mining
- Ad optimization

Want to learn more about Hive?



Contact info

@mark_grover

github.com/markgrover

linkedin.com/in/grovermark

mgrover@cloudera.com