cloudera
Ask Bigger Questions

Introduction to Apache Hive

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

Agenda

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

Agenda

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

Hive

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

Agenda

- What is Hive?
- Why use Hive?
- Hive features
- Hive architecture
- 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
- 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

Other Hive features

- Partitioning
- Sampling
- Bucketing
- Various kinds of optimized joins
- Integration with HBase and other storage handlers

Connecting to Hive

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

Agenda

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

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

Agenda

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

Contact info

@mark_grover

github.com/markgrover

linkedin.com/in/grovermark

mgrover@cloudera.com