SQL-Like Languages

CISC-525

Phil Grim

HARRISBURG UNIVERSITY
OF SCIENCE AND TECHNOLOGY

Overview

- Storage Formats
- Hive
- Pig
- Drill

Storage Formats

- Text File
- Sequence File
- RC File
- ORC File
- Parquet File

Text Files

- Simple human readable files
- Uncompressed*
- Delimited
  - CSV
  - TSV
- Good for importing and exporting data

Sequence Files

- Hadoop's native format
- Compressed
- Binary
- Can be split

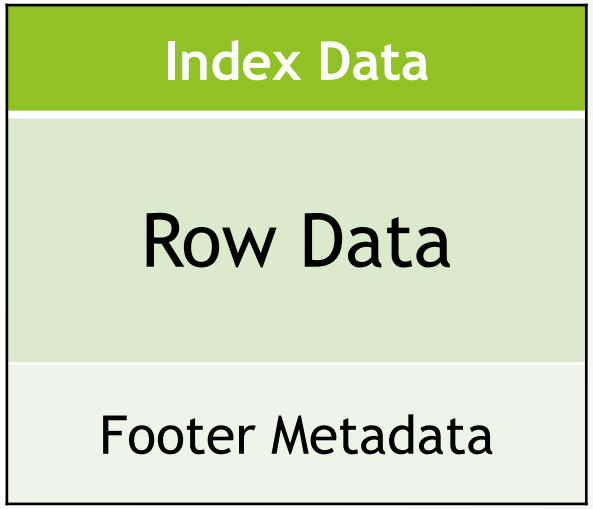
RC File Format

- Columnar store format
- Optimizes query performance
- Stores columns in row groups
- Cannot be split other than on block boundaries

Text File	RCFile
1, One, Alpha, First 2, Two, Beta, Second 3, Three, Gamma, Third 4, Four, Delta, Fourth	1,2,3,4 One, Two, Three, Four Alpha, Beta, Gamma, Delta First, Second, Third, Fourth

ORC File Format

- Optimized RCFile
- Compressed
- Adds header and footer
  - Indexing
  - Metadata
- Can be natively split



Parquet File Format

- Alternative to ORC
- Claims 10x performance over sequence files
- Compressed
- Shared across many tools

What is Hive?

- A data warehouse framework for interacting with the Hadoop ecosystem
- Provides a language called Hive Query Language (HQL)
  - Structure and syntax similar to Structured Query Language (SQL)
  - Lowers the learning curve for developers to leverage Hadoop
- Many ways to access data
  - Interactively through the Hive shell
  - JDBC/ODBC
  - From applications using Thrift
  - From other Hadoop ecosystem components using HCatalog

What is Hive not?

- An RDBMS
- Highly interactive
- An Online Transaction Processing (OLTP) system

Hive Architecture

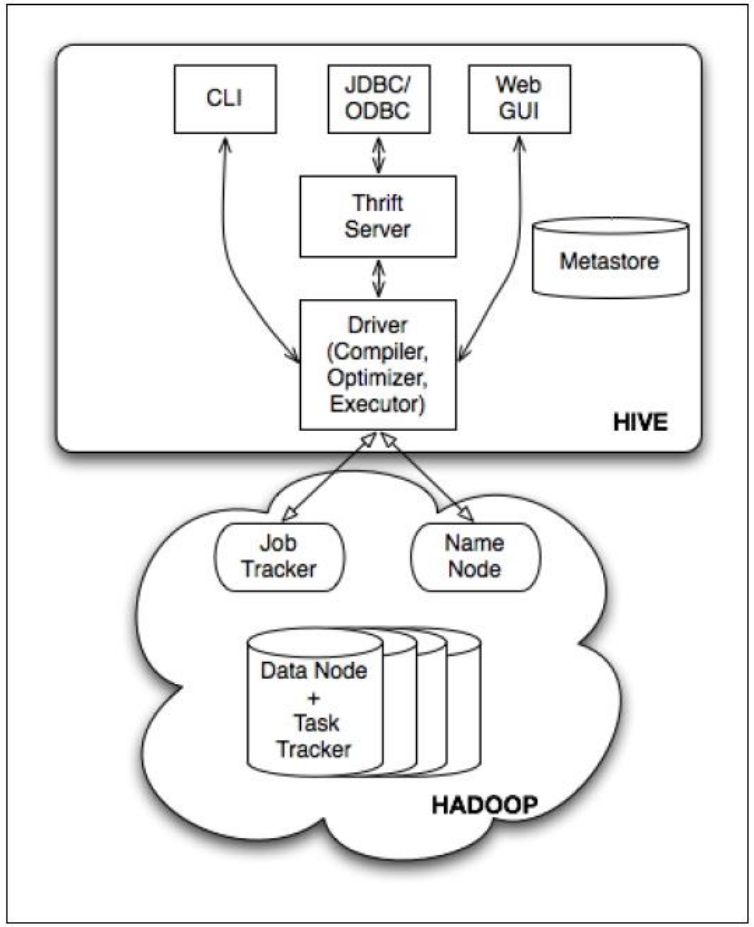
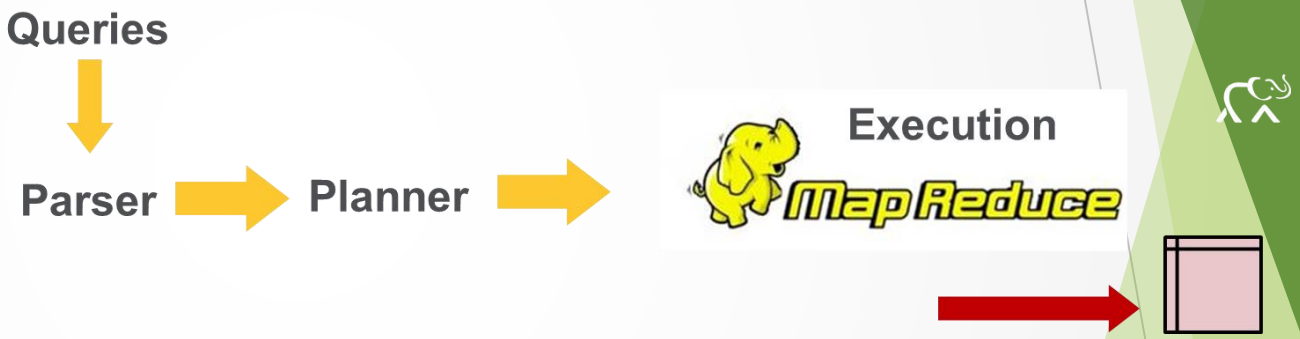


Figure 1: Hive Architecture



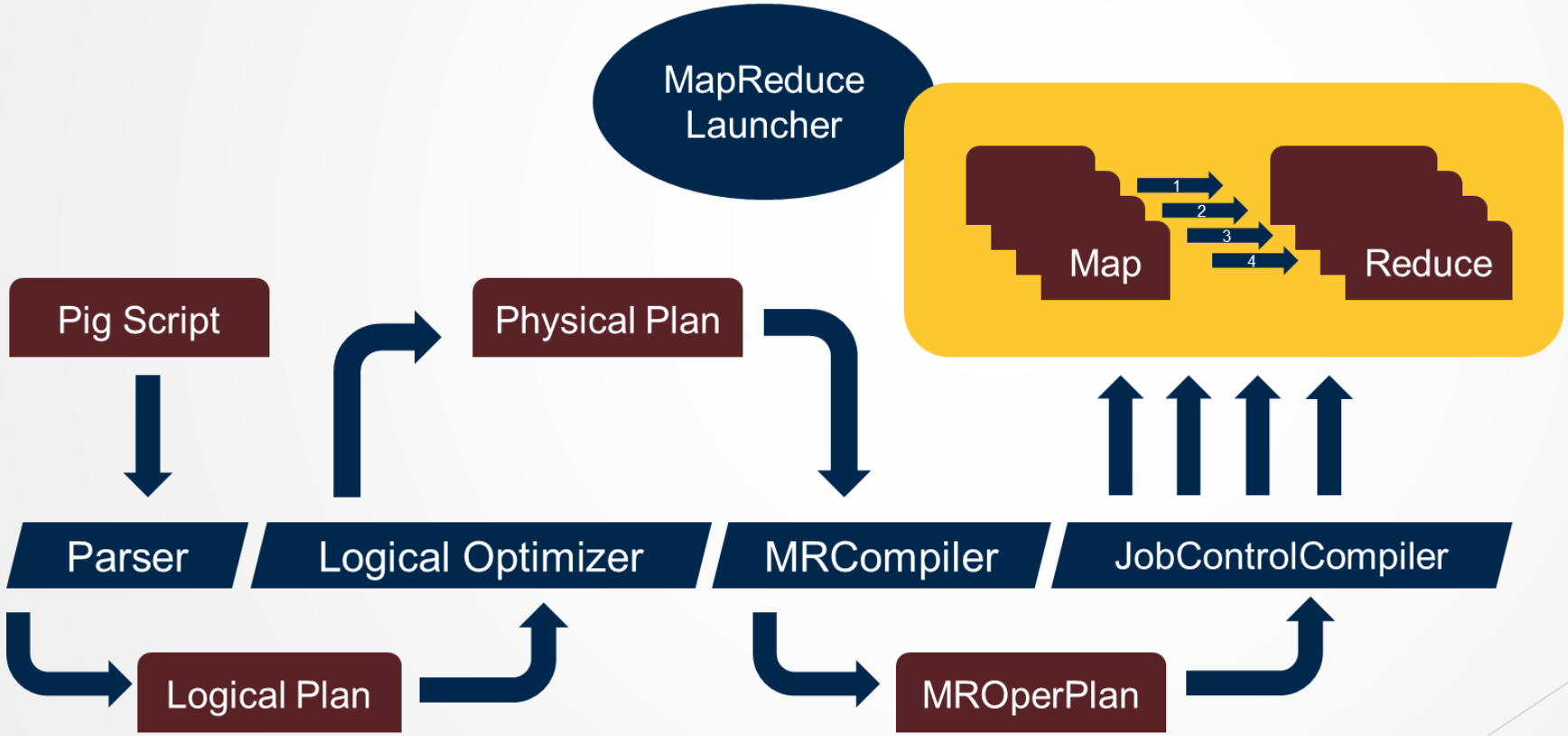
What is Pig?

- A data analysis framework for interacting with the Hadoop ecosystem
- Provides a scripting language called Pig Latin
  - Simple syntax with some features in common with SQL
  - A data flow language rather than a query language
- Executes Pig commands using MapReduce
  - Interactively through the Grunt shell
  - Batch mode using script files
- Able to interact with the HCatalog to access data stored by Hive

What is Pig not?

- An RDBMS
- A true programming language
- A query language

Pig Architecture



Pig Latin

Example:

```
A = LOAD 'student' AS (name:chararray, age:int, gpa:float);
X = FOREACH A GENERATE name,$2;
DUMP X;
(John,4.0F)
(Mary,3.8F)
(Bill,3.9F)
(Joe,3.8F)
```

Pig Latin

```
DUMP A;
(John,18,4.0F)
(Mary,19,3.8F)
(Bill,20,3.9F)
(Joe,18,3.8F)

B = GROUP A BY age;

DUMP B;
(18,{(John,18,4.0F),(Joe,18,3.8F)})
(19,{(Mary,19,3.8F)})
(20,{(Bill,20,3.9F)})
```

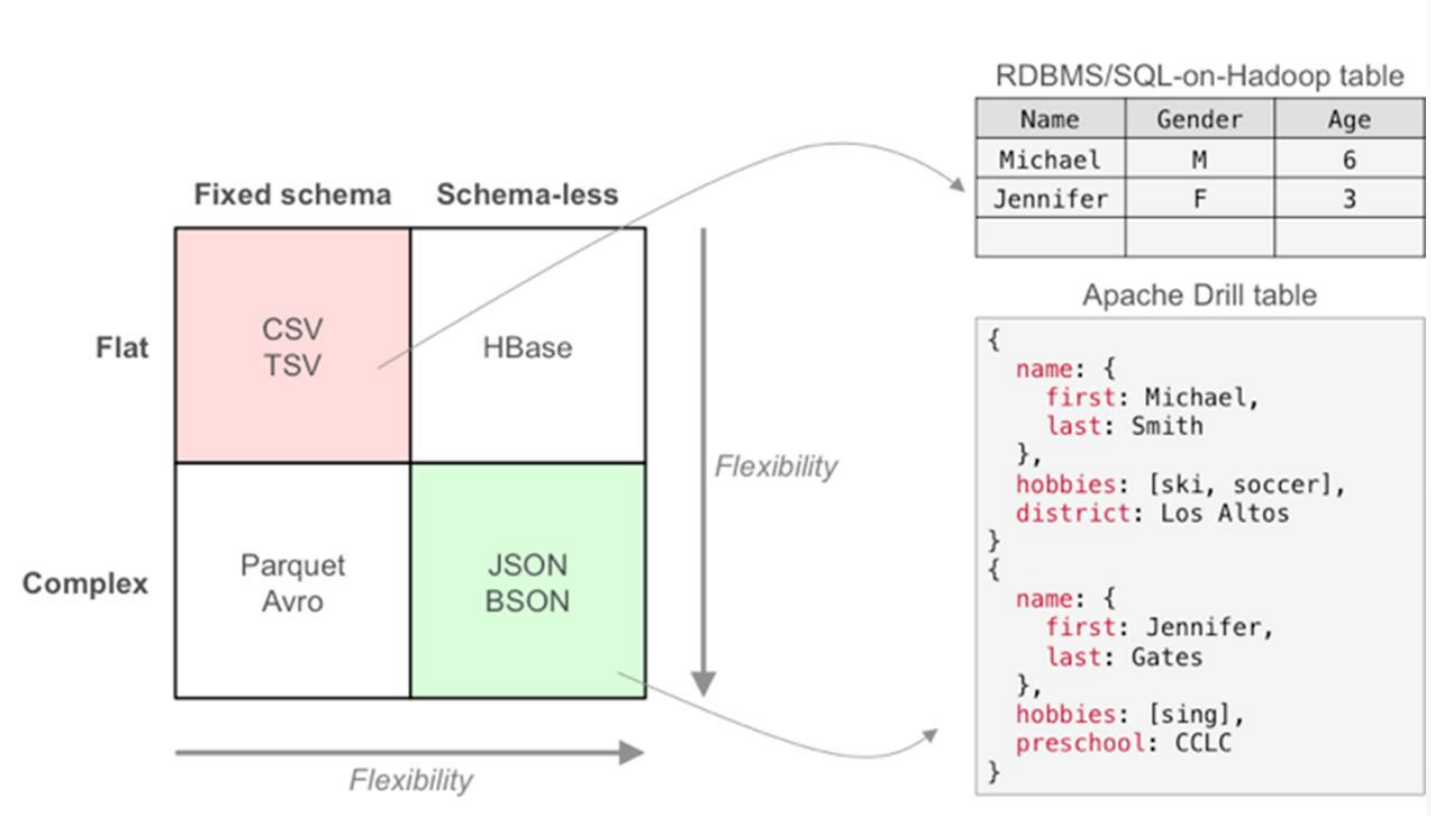
What is Drill?

- Open-source low-latency SQL query engine
- Schema-less
- Works well with semi-structured and self-describing data
  - JSON
  - Nested/Complex types
  - Compatible with SQL databases and with Hive and HBase
- Provides JDBC connectivity

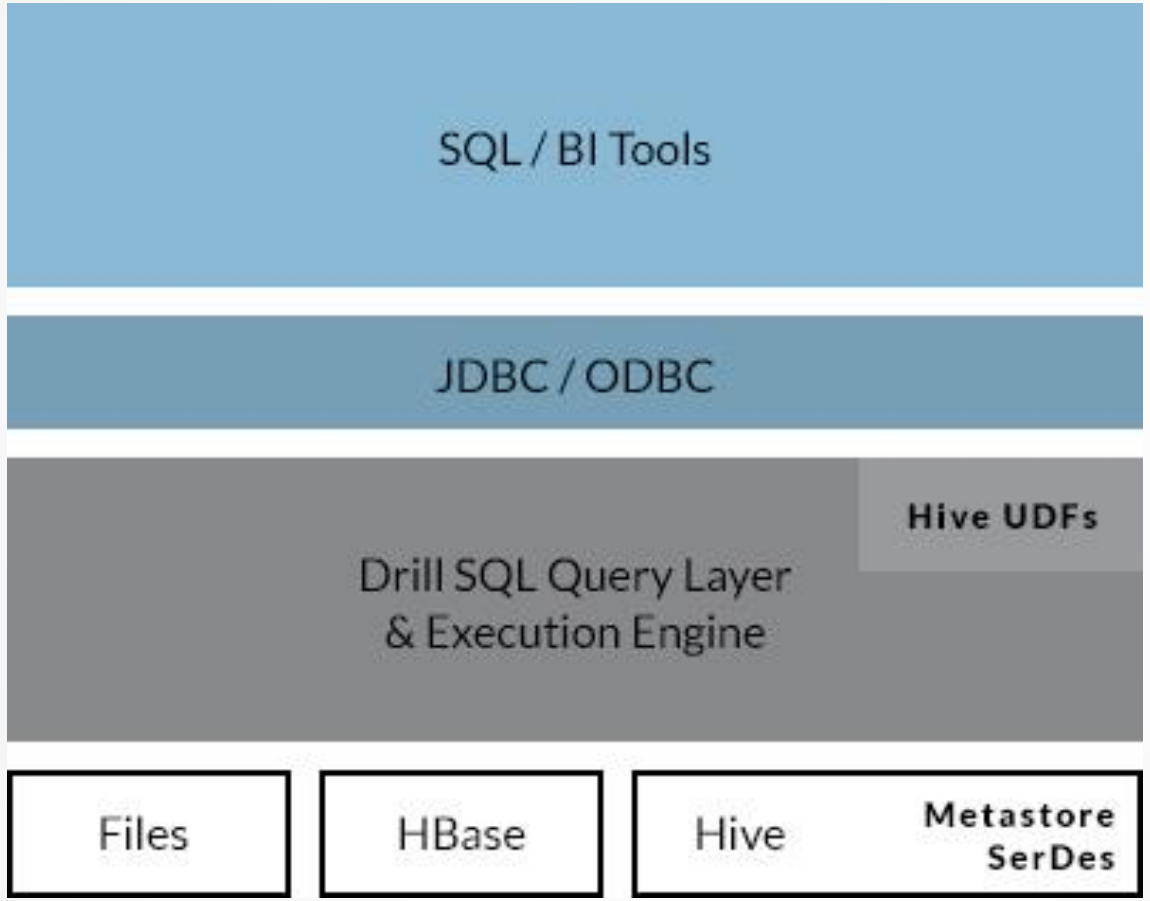
What is Drill Not?

- An RDBMS
- A Front End to MapReduce

Drill Model


Fixed schema
Schema-less
Flat
Complex
CSV
TSV
HBase
Parquet
Avro
JSON
BSON
Flexibility
Flexibility
RDBMS/SQL-on-Hadoop table
Name
Gender
Age
Michael
M
6
Jennifer
F
3
Apache Drill table
{
name: {
first: Michael,
last: Smith
},
hobbies: [ski, soccer],
district: Los Altos
}
{
name: {
first: Jennifer,
last: Gates
},
hobbies: [sing],
preschool: CCLC
}

Drill Architecture



Continued Reading

Hive Website

http://hive.apache.org

Pig Website

http://pig.apache.org

Drill Website

http://drill.apache.org