Storage and the eco-system

Ahmad Alkilani
www.pluralsight.com

pluralsight
hardcore developer training

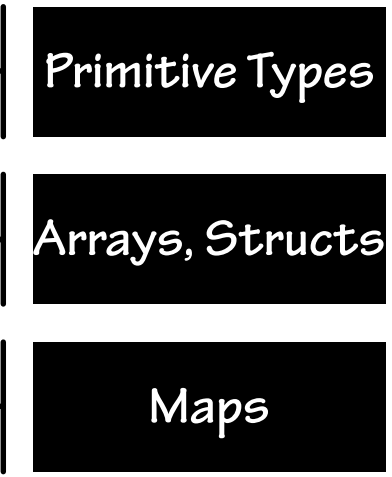
Outline

- **Create Table**
  - Storage Formats
  - Serializers/Deserializers (SerDes)
- **File Format Examples**
- **HCatalog**
- **Eco-System Projects**
  - Sqoop
  - DistCp
  - Others worth mentioning
- **Course Resources**

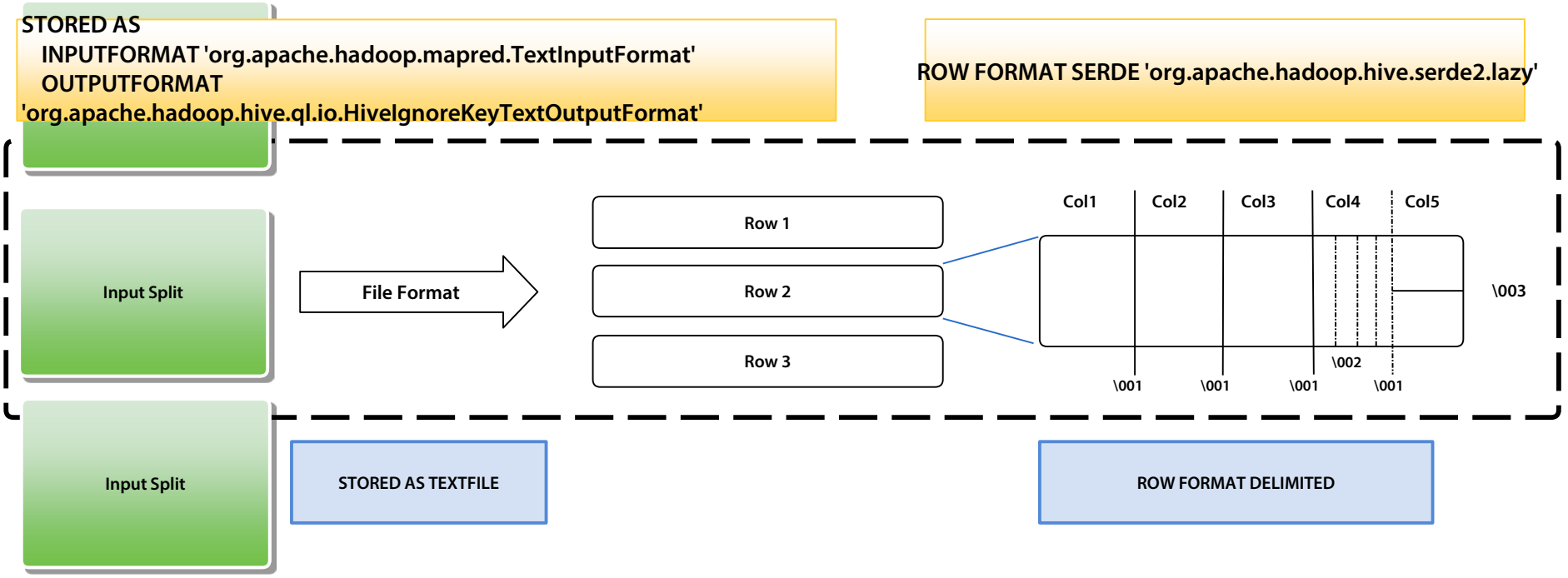
Create Table In-Depth

```
CREATE TABLE ……
[ROW FORMAT row_format]
[STORED AS file_format]
```

```
CREATE TABLE ……
ROW FORMAT DELIMITED
    FIELDS TERMINATED BY '\001'
    COLLECTION ITEMS TERMINATED BY
    '\002'
    MAP KEYS TERMINATED BY '\003'
    LINES TERMINATED BY '\n'
STORED AS TEXTFILE;
```



Create Table In-Depth



File Formats

- **Text File**

```
CREATE TABLE t1 (a INT, b STRING, c STRING)
ROW FORMAT DELIMITED
STORED AS TEXTFILE;
```

- **Sequence File**

```
CREATE TABLE t1 (a INT, b STRING, c STRING)
STORED AS SEQUENCEFILE;
```

File Formats (2)

- RCFile

```
CREATE TABLE t1 (a INT, b STRING, c STRING)
ROW FORMAT SERDE
  'org.apache.hadoop.hive.serde2.columnar.ColumnarSerDe'
STORED AS
  INPUTFORMAT 'org.apache.hadoop.hive.ql.io.RCFileInputFormat'
  OUTPUTFORMAT 'org.apache.hadoop.hive.ql.io.RCFileOutputFormat';
```

- ORCFile

```
CREATE TABLE t1 (a INT, b STRING, c STRING)
STORED AS orc
TBLPROPERTIES (
"orc.compress"="SNAPPY"
,"orc.create.index"="true"
);
```

HCatalog

- **Set of Interfaces (APIs) and metadata service**
- **Doesn't re-invent a meta-store that already works well**
  - Sits on top of Hive's meta-store
- **Centralizes metadata services for the Hadoop eco-system**
- **Gives tools like Apache Pig and MapReduce an abstraction layer**
  - Databases, tables, partitions etc..
  - File storage format and location

HCatalog (2)

- Access through the CLI

```
hcat -e "describe pluralsight.movies;"
hcat -e "show tables from pluralsight;"
```

- REST API

- Usage within PIG

```
A = load 'pluralsight.movies' using HCatLoader();
B = filter A by name= 'Despicable Me';
…
```

Sqoop (SQL-to-Hadoop)

- Transfer data to and from relational databases
- Anything with a JDBC driver
- Specific optimization working with MySQL

```
sqoop --connect jdbc:mysql://someURI --table flights --hive-import
```

- Parallel Import
  - Launches multiple mappers, each with a subset of the query
    - Default 4 map tasks
  - By default Sqoop identifies the primary key if present

```
sqoop import --connect <connect-str> --passwordFile ${user.home}/.password \
--query 'SELECT a.x, b.y FROM A JOIN B on (a.id == b.id) WHERE $CONDITIONS' \
--split-by a.userid --target-dir /some/location/on/hdfs/mydata
[-m|--num-mappers number]
```

Sqoop (2)

- **Incremental Imports**
  - Append
    - Increasing Row IDs
  - Last Modified
    - Records that are updated

  --check-column to specify which column to use

- **Learn more:**

  http://sqoop.apache.org/docs/1.4.4/SqoopUserGuide.html

DistCP Version 2

- **Tool used to copy large amounts of data**
  - Within a cluster
  - To/From different clusters

- **Uses MapReduce underneath to parallelize and achieve fault tolerance**

```
hadoop distcp2 hdfs://nn1:8020/my/data
    hdfs://nn2:8020/copy/of/my/
```

- **Options to Update and Overwrite**
  - -update
  - -overwrite

Other Eco-System Projects

- **Scalding**
  - Based on Scala
  - Created by and extensively at Twitter
- **Apache Mahout**

- **Impala**
  - Fairly new
  - Created by Cloudera to address “Real-time” querying for Hadoop
  - Not as mature as Hive
- **Apache Drill**
  - Based on Google’s Dremel
  - Interactive
  - Nested data
  - Potentially an execution for Hive

- **Storm**
  - Streaming

References and Resources

- **Pluralsight!**
- **Wiki**
  - http://hadoop.apache.org/
- **Apache Hive Wiki and Language Manual**
  - https://cwiki.apache.org/confluence/display/Hive/Home
  - https://cwiki.apache.org/confluence/display/Hive/LanguageManual
- **Programming Hive, O'REILLY**
- **Hadoop The Definitive Guide, O'REILLY**
- **Sqoop**
  - http://sqoop.apache.org/
- **DistCP Version 2**
  - http://hadoop.apache.org/docs/r1.2.0/distcp2.html
- **Apache Drill**
  - http://incubator.apache.org/drill/
- **Hortonworks Docs**
  - http://docs.hortonworks.com

References and Resources

- **Hadoop Eclipse Plug-in**
  - http://wiki.apache.org/hadoop/EclipsePlugIn
- **Develop CDH Applications with Maven and Eclipse**
  - http://blog.cloudera.com/blog/2012/08/developing-cdh-applications-with-maven-and-eclipse/
- **Apache BigTop**
  - http://apachebigtop.pbworks.com/w/page/48434924/FrontPage

Summary

- **Create Table**
  - Storage Formats
  - Serializers/Deserializers (SerDes)
- **File Format Examples**
- **HCatalog**
- **Eco-System Projects**
  - Sqoop
  - DistCp
  - Others worth mentioning
- **Course Resources**