## OM ASHISH MISHRA

1. Process “xml\_data.txt” using Hive or Pig then export the xml contents to MYSQL database using sqoop. Create the table schema to store all the contents assume types as per data content.

METHOD 1:

XML TO HIVE

add jar /home/cloudera/hivexmlserde-1.0.0.0.jar;

CREATE TABLE xml\_data

(id string,

author string,

title string,

genre string,

price float,

publish\_date date,

description string)

ROW FORMAT SERDE 'com.ibm.spss.hive.serde2.xml.XmlSerDe'

with serdeproperties(

"column.xpath.id"="/book/id/text()",

"column.xpath.author"="/book/author/text()",

"column.xpath.title"="/book/title/text()",

"column.xpath.genre"="/book/genre/text()",

"column.xpath.price"="/book/price/text()",

"column.xpath.publish\_date"="/book/publish\_date/text()",

"column.xpath.description"="/book/description/text()")

stored as inputformat 'com.ibm.spss.hive.serde2.xml.XmlInputFormat'

outputformat 'org.apache.hadoop.hive.ql.io.IgnoreKeyTextOutputFormat'

tblproperties("xmlinput.start"="<book>","xmlinput.end"="</book>");

LOAD DATA LOCAL INPATH 'xml\_data.txt' INTO TABLE xml\_data;

SELECT distinct genre FROM xml\_data;

HIVE TO MYSQL

Mysql –u root –p (cloudera)

Create table xml\_data(

id varchar(100),

author varchar(100),

title varchar(100),

genre varchar(100),

price decimal(14,2),

publish\_date date,

description varchar(100));

exit;

sqoop export --connect “jdbc://localhost:3306/retail.db” --username root --password cloudera --table xml\_data --export-dir ‘/user/hive/warehouse/aspire.db/xml\_data/xml\_data.json’ --input-fields-terminated-by ‘,’ --input-fields-terminated-by ‘\n’ --num mappers 2 --outdir java\_files

mysql –u root –p (cloudera)

use retail\_db;

select \* from xml\_data;

METHOD 2:

HADOOP TO PIG

grunt> register piggybank-0.15.0.jar;

grunt> A1 = LOAD 'xml\_data.txt' using org.apache.pig.piggybank.storage.XMLLoader('book') AS (x:chararray);

grunt> define XPath org.apache.pig.piggybank.evaluation.xml.XPath();

grunt> B1 = FOREACH A1 GENERATE XPath(x,'book/id'), XPath(x,'book/author'), XPath(x,'book/title'), XPath(x,'book/genre'), XPath(x,'book/publish\_date'), XPath(x,'book/description');

grunt> STORE B1 INTO 'ans2' using PigStorage(',');

PIG TO HIVE

use aspire;

CREATE TABLE xml\_data1(id string,author string,title string,genre string,publish\_date string,description string)

ROW FORMAT DELIMITED FIELDS TERMINATED BY ',';

hive> LOAD DATA LOCAL INPATH '/home/cloudera/ans2' INTO table xml\_data1;

hive> SELECT \* FROM xml\_data1;

HIVE TO MYSQL

Mysql –u root –p (cloudera)

Create table xml\_data(

id varchar(100),

author varchar(100),

title varchar(100),

genre varchar(100),

price decimal(14,2),

publish\_date date,

description varchar(100));

exit;

sqoop export --connect "jdbc:mysql://localhost:3306/retail\_db" --username root --password cloudera --table xml\_data --export-dir '/user/hive/warehouse/aspire.db/xml\_data1/xml\_data.txt' --input-fields-terminated-by ',' --input-lines-terminated-by '\n' --num-mappers 2 --outdir java\_files

mysql –u root –p (cloudera)

use retail\_db;

select \* from xml\_data;

1. Process “world\_bank.json” using Hive and export the contents to MYSQL database using sqoop. Create the table schema to store all the contents assume types as per data content.

METHOD 1 :

JSON TO HIVE

add jar /home/cloudera/json-serde-1.3.7-jar-with-dependencies.jar;

create table json\_serde

(uid struct<oid:string>,approvalfy int,board\_approval\_month string,boardapprovaldate string,borrower string,country\_namecode string,countrycode string,countryname string,countryshortname string,docty string,envassesmentcategorycode string,grantamt bigint,ibrdcommamt int,id string,idacommamt int,impagency string,lendinginstr string,lendinginstrtype string,lendprojectcost bigint,majorsector\_percent struct<Name:string,Percent:int>,mjsector\_namecode struct<name:string,code:string>,mjthemecode string,prodline string,prodlinetext string,productlinetype string,project\_name string,projectdocs struct<DocTypeDesc:string,DocType:string,EntityID:string,DocURL:string,DocDate:string>,projectfinancialtype string,projectstatusdisplay string,regionname string,sector struct<Name:string>,sector1 struct<Name:string,Percent:int>,sector2 struct<Name:string,Percent:int>,sector\_namecode struct<name:string,code:string>,sectorcode string,source string,status string,supplementprojectflg string,theme1 struct<Name:string,Percent:int>,theme\_namecode struct<name:string,code:string>,themecode string,totalamt int,totalcommamt bigint,url string)

ROW FORMAT SERDE 'org.openx.data.jsonserde.JsonSerDe';

LOAD DATA LOCAL INPATH '/home/cloudera/world\_bank.json' INTO TABLE json\_serde;select \* from json\_serde;describe json\_serde;

ERROR SHOWN:-

Failed with exception java.io.IOException:org.apache.hadoop.hive.ql.metadata.HiveException: java.lang.ClassCastException: org.openx.data.jsonserde.json.JSONObject cannot be cast to java.lang.Integer

Time taken: 0.872 seconds

WARN: The method class org.apache.commons.logging.impl.SLF4JLogFactory#release() was invoked.

WARN: Please see http://www.slf4j.org/codes.html#release for an explanation.

desc json\_serde;

ERROR SHOWN:-

FAILED: Execution Error, return code 1 from org.apache.hadoop.hive.ql.exec.DDLTask. java.lang.ClassNotFoundException Class org.openx.data.jsonserde.JsonSerDe not found

HIVE TO MYSQL

Mysql –u root –p (cloudera)

Create table json\_serde(uid\_oid varchar(100),approvalfy integer,board\_approval\_month varchar(100),boardapprovaldate varchar(100),borrower varchar(100),country\_namecode varchar(100),countrycode varchar(100),countryname varchar(100),countryshortname varchar(100),docty varchar(100),envassesmentcategorycode varchar(100),grantamt int,ibrdcommamt int,id varchar(100),idacommant int,impagency varchar(100),lendinginstr varchar(100),lendinginstrtype varchar(100),lendprojectcost int,majorsector\_percent\_Name varchar(100),majorsector\_percent\_Percent int,mjsector\_namecode\_name varchar(100),mjsector\_namecode\_code varchar(100),mjthemecode varchar(100),prodline varchar(100),prodlinetext varchar(100),productlinetype varchar(100),project\_name varchar(100),projectdocs\_DoctypeDesc varchar(100),projectdocs\_DocType varchar(100),projectdocs\_EntityID varchar(100),projectdocs\_DocURL varchar(100),projectdocs\_DocDate varchar(100),projectfinancialtype varchar(100),projectstatusdisplay varchar(100),regionname varchar(100),sector\_Name varchar(100),sector1\_Name varchar(100),sector1\_Percent integer,sector2\_Name varchar(100),sector2\_Percent integer,sector\_namecode\_name varchar(100),sector\_name\_code varchar(100),sectorcode varchar(100),source varchar(100),status varchar(100),supplementprojectflg varchar(100),theme1\_Name varchar(100),theme1\_Percent integer,theme\_namecode\_name varchar(100),theme\_namecode\_code varchar(100),themecode varchar(100),totalamt integer, totalcommamt integer,url varchar(100));

exit;

sqoop export --connect “jdbc://localhost:3306/retail.db” --username root --password cloudera --table json\_serde --export-dir ‘/user/hive/warehouse/aspire.db/json\_serde/world\_bank.json’ --input-fields-terminated-by ‘,’ --input-fields-terminated-by ‘\n’ --num mappers 2 --outdir java\_files

mysql –u root –p (cloudera)

use retail\_db;

select \* from json\_serde;

METHOD 2:

HADOOP TO PIG