

Data Model Reference

Tables Imported from MySQL

The following depicts the structure of the MySQL tables imported into HDFS using Sqoop. The primary key column from the database, if any, is denoted by bold text:

customers: 201,375 records (imported to `/dualcore/customers`)

Index	Field	Description	Example
0	cust_id	Customer ID	1846532
1	fname	First name	Sam
2	lname	Last name	Jones
3	address	Address of residence	456 Clue Road
4	city	City	Silicon Sands
5	state	State	CA
6	zipcode	Postal code	94306

employees: 61,712 records (imported to `/dualcore/employees` and later used as an external table in Hive)

Index	Field	Description	Example
0	emp_id	Employee ID	BR5331404
1	fname	First name	Betty
2	lname	Last name	Richardson
3	address	Address of residence	123 Shady Lane
4	city	City	Anytown
5	state	State	CA
6	zipcode	Postal Code	90210
7	job_title	Employee's job title	Vice President
8	email	e-mail address	br5331404@example.com
9	active	Is actively employed?	Y
10	salary	Annual pay (in dollars)	136900

orders: 1,662,951 records (imported to `/dualcore/orders`)

Index	Field	Description	Example
0	order_id	Order ID	3213254
1	cust_id	Customer ID	1846532
2	order_date	Date/time of order	2013-05-31 16:59:34

order_details: 3,333,244 records (imported to /dualcore/order_details)

Index	Field	Description	Example
0	order_id	Order ID	3213254
1	prod_id	Product ID	1754836

products: 1,114 records (imported to /dualcore/products)

Index	Field	Description	Example
0	prod_id	Product ID	1273641
1	brand	Brand name	Foocorp
2	name	Name of product	4-port USB Hub
3	price	Retail sales price, in cents	1999
4	cost	Wholesale cost, in cents	1463
5	shipping_wt	Shipping weight (in pounds)	1

suppliers: 66 records (imported to /dualcore/suppliers)

Index	Field	Description	Example
0	supp_id	Supplier ID	1000
1	fname	First name	ACME Inc.
2	lname	Last name	Sally Jones
3	address	Address of office	123 Oak Street
4	city	City	New Athens
5	state	State	IL
6	zipcode	Postal code	62264
7	phone	Office phone number	(618) 555-5914

Hive Tables

The following is a record count for Hive tables that are created or queried during the hands-on exercises. Use the `DESCRIBE tablename` command in Hive to see the table structure.

Table Name	Record Count
cart_items	33,812
cart_orders	12,955
cart_shipping	12,955
cart_zipcodes	12,955
checkout_sessions	12,955
customers	201,375
employees	61,712
order_details	3,333,244
orders	1,662,951
products	1,114
ratings	21,997
web_logs	412,860

Other Data Added to HDFS

The following describes the structure of other important data sets added to HDFS.

Combined Ad Campaign Data: (788,952 records total), stored in two directories:

- /dualcore/ad_data1 (438,389 records)
- /dualcore/ad_data2 (350,563 records).

Index	Field	Description	Example
0	campaign_id	Uniquely identifies our ad	A3
1	date	Date of ad display	05/23/2013
2	time	Time of ad display	15:39:26
3	keyword	Keyword that triggered ad	tablet
4	display_site	Domain where ad shown	news.example.com
5	placement	Location of ad on Web page	INLINE
6	was_clicked	Whether ad was clicked	1
7	cpc	Cost per click, in cents	106

access.log: 412,860 records (uploaded to /dualcore/access.log)

This file is used to populate the `web_logs` table in Hive. Note that the RFC 931 and Username fields are seldom populated in log files for modern public Web sites and are ignored in our `RegexSerDe`.

Index	Field / Description	Example
0	IP address	192.168.1.15
1	RFC 931 (Ident)	-
2	Username	-
3	Date/Time	[22/May/2013:15:01:46 -0800]
4	Request	"GET /foo?bar=1 HTTP/1.1"
5	Status code	200
6	Bytes transferred	762
7	Referer	"http://dualcore.com/"
8	User agent (browser)	"Mozilla/4.0 [en] (WinNT; I) "
9	Cookie (session ID)	"SESSION=8763723145"