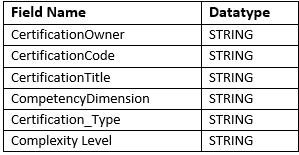
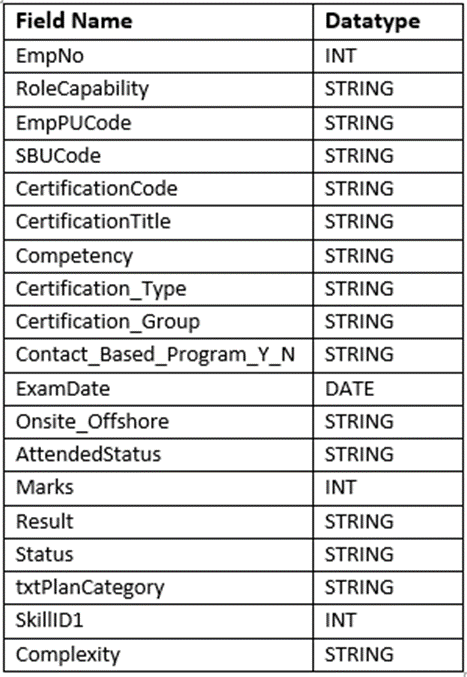
**Dataset:**

Data source : The required data sets are provided in the below links

[CertificationDetails](https://lex.infosysapps.com/content-store/Infosys/Infosys_Ltd/Public/lex_auth_012797658897768448643/web-hosted/assets/CertificationDetails.csv)



[CDPCompetionStatus](https://lex.infosysapps.com/content-store/Infosys/Infosys_Ltd/Public/lex_auth_012797658897768448643/web-hosted/assets/CDPCompletionStatus.csv)



**Task 1:** **Working with HDFS & Sqoop**

1. In MySQL, create database called **CertificationDB\_empno**. In this database, create two new tables called **CertificationDetails** and **CDPCompletionStatus**with the columns mentioned in the schema above

**Mysql tables:**

mysql> create table CertificationDetails(CertificationOwner Varchar(50),CertificationCode Varchar(30),CertificationTitle Varchar(100), CompetencyDimension Varchar(20),Certification\_Type Varchar(20),ComplexityLevel Varchar(20));

mysql> create table CDPCompletionStatus(Empno Int(10),RoleCapability Varchar(50),EmpPUCode Varchar(20),SBUCode Varchar(20),CertificationCode Varchar(30), CertificationTitle Varchar(100),competency Varchar(20),Certification\_Type Varchar(20),Certification\_Group Varchar(20),Contact\_Based\_Program\_Y\_N Varchar(5), ExamDate Date,Onsite\_Offshore Varchar(20),AttendedStatus Varchar(20),Marks Int,Result Varchar(20),Status Varchar(20),txtPlanCategory Varchar(20), SkillID1 int,Complexity Varchar(20));

Query OK, 0 rows affected (0.04 sec)

2. Create a new HDFS directory called **HDFS\_empno**. Move the two sample data csv files to this HDFS directory using HDFS commands

hadoop fs -put CertificationDetails.csv /user/hduser/HDFS\_empno\_lex/

hadoop fs -put CDPCompletionStatus.csv /user/hduser/HDFS\_empno\_lex/

3. Transfer the data present in HDFS\_empno directory to the 2 tables created in MySQL

sqoop export --connect jdbc:mysql://localhost/Lex\_CertificationDB\_empno --username root --password root --table CertificationDetails

--export-dir /user/hduser/HDFS\_empno\_lex/CertificationDetails.csv

-m 1 --fields-terminated-by ',' --lines-terminated-by '\n' **–batch**

sqoop export --connect jdbc:mysql://localhost/Lex\_CertificationDB\_empno --username root --password root --table CDPCompletionStatus

--export-dir /user/hduser/HDFS\_empno\_lex/CDPCompletionStatus.csv

--fields-terminated-by ',' --lines-terminated-by '\n' -m 2 **–direct**

4. Transfer the data from both the MySQL tables to a new destination HDFS directory named Certification\_empno

sqoop **import-all-tables** --connect jdbc:mysql://localhost/Lex\_CertificationDB\_empno --username root --password root --warehouse-dir /user/hduser/Certification\_empno\_lex -m 1

5. Copy the files from **Certification\_empno**directory to **Backup\_empno** directory

hadoop fs -cp /user/hduser/Certification\_empno\_lex /user/hduser/Backup\_empno\_lex

**Task 2: Data Analysis using Hive**

1. Create two Hive tables for the provided datasets

hive> create database certification\_empno\_lex

comment 'LEx workouts'

with dbproperties('Created by'='Inceptez','Created on'='10-11-2020');

hive (certification\_empno\_lex)> create table certificationdetails(certificationowner string,certificationcode string,certificationtitle string,competencydimension string, complexitylevel string)

**partitioned by(certification\_type string)**

**ROW FORMAT SERDE 'org.apache.hadoop.hive.serde2.OpenCSVSerde'**

**WITH SERDEPROPERTIES (**

**"seperatorChar"=",",**

**"quoteChar"="\"");**

hive (certification\_empno\_lex)> LOAD DATA LOCAL INPATH '/home/hduser/hiveusecase/Hive\_datasets/CertificationDetails.csv'

OVERWRITE INTO TABLE certificationdetails ;

hive (certification\_empno\_lex)> create table CDPcompletionstatus(empno INT,rolecapabilty string,emppucode string,sbucode string,certificationcode string,certificationtitle string,compentency string,certification\_type string,certification\_group string,contact\_bases\_program\_Y\_N string,examdate DATE,onsite\_offshore string,attendedstatus string,marks INT,result string,status string,txtplancategory string,skillid1 INT,complexity string)

**ROW FORMAT SERDE 'org.apache.hadoop.hive.serde2.OpenCSVSerde'**

**WITH SERDEPROPERTIES (**

**"seperatorChar"=",",**

**"quoteChar"="\"");**

LOAD DATA LOCAL INPATH '/home/hduser/hiveusecase/Hive\_datasets/CDPCompletionStatus.csv' OVERWRITE INTO TABLE CDPCompletionStatus;

2. List the count of employees qualified for various certifications

select count(\*)as Empcount from CDPcompletionstatus where result = 'Qualified';

3. List the certifications available under a specific SBUCode

select sbucode,certificationcode,certificationtitle from CDPcompletionstatus group by sbucode,certificationcode,certificationtitle;

4.List the count of employees attempted a particular certification from a specific unit (column: EmpPUCode)

select emppucode, certificationtitle,count(empno) from CDPcompletionstatus group by emppucode,certificationtitle;

hive (certification\_empno\_lex)> LOAD DATA LOCAL INPATH '/home/hduser/hiveusecase/Hive\_datasets/CertificationDetails.csv'

OVERWRITE INTO TABLE certificationdetails

**PARTITION(certification\_type='External')**

;

5. Split the CertificationDetails table into two partitions using dynamic partitioning based on the values of Certification\_Type column

Staging table:

create table certificationdetailsstage(certificationowner string,certificationcode string,certificationtitle string,competencydimension string, certification\_type string,complexitylevel string)

ROW FORMAT SERDE 'org.apache.hadoop.hive.serde2.OpenCSVSerde'

WITH SERDEPROPERTIES (

"seperatorChar"=",",

"quoteChar"="\"");

LOAD DATA LOCAL INPATH '/home/hduser/hiveusecase/Hive\_datasets/CertificationDetails.csv'

OVERWRITE INTO TABLE certificationdetailsstage ;

hive (certification\_empno\_lex)> Insert overwrite table certificationdetails **partition(certification\_type)**

select certificationowner,certificationcode,certificationtitle,competencydimension,complexitylevel,**certification\_type** from certificationdetailsstage;

hive (certification\_empno\_lex)> Insert overwrite table certificationdetails **partition(certification\_type=’External’)**

select certificationowner,certificationcode,certificationtitle,competencydimension,complexitylevel,certification\_type from certificationdetailsstage where **certification\_Type=’External’;**

hive (certification\_empno\_lex)> Insert overwrite table certificationdetails **partition(certification\_type=’Internal’)**

select certificationowner,certificationcode,certificationtitle,competencydimension,complexitylevel,certification\_type from certificationdetailsstage **where certification\_Type=’Internal’**;

create table BucketCDPcompletionstatus(empno INT,rolecapabilty string,emppucode string,sbucode string,certificationcode string,certificationtitle string,compentency string,certification\_type string,certification\_group string,contact\_bases\_program\_Y\_N string,examdate DATE,onsite\_offshore string,attendedstatus string,marks INT, result string,status string,txtplancategory string,skillid1 INT,complexity string)

**CLUSTERED BY(marks) INTO 5 buckets**

ROW FORMAT SERDE 'org.apache.hadoop.hive.serde2.OpenCSVSerde'

WITH SERDEPROPERTIES (

"seperatorChar"=",",

"quoteChar"="\"")

Insert overwrite table BucketCDPCompletionstatus

select \* from CDPCompletionstatus;

**Task 3: Data migration using Sqoop**

1. Transfer the data from the table CertificationDetails in MySQL to HDFS directory Sqoop\_data\_empno using Sqoop. Output file should contain only Advanced level certification

sqoop import --connect jdbc:mysql://localhost/Lex\_CertificationDB\_empno --username root --password root --table CertificationDetails

--where "complexitylevel='Advanced'"

--target-dir /user/hduser/sqoop\_data\_empno --delete-target-dir

--fields-terminated-by ',' --lines-terminated-by '\n' -m 1