

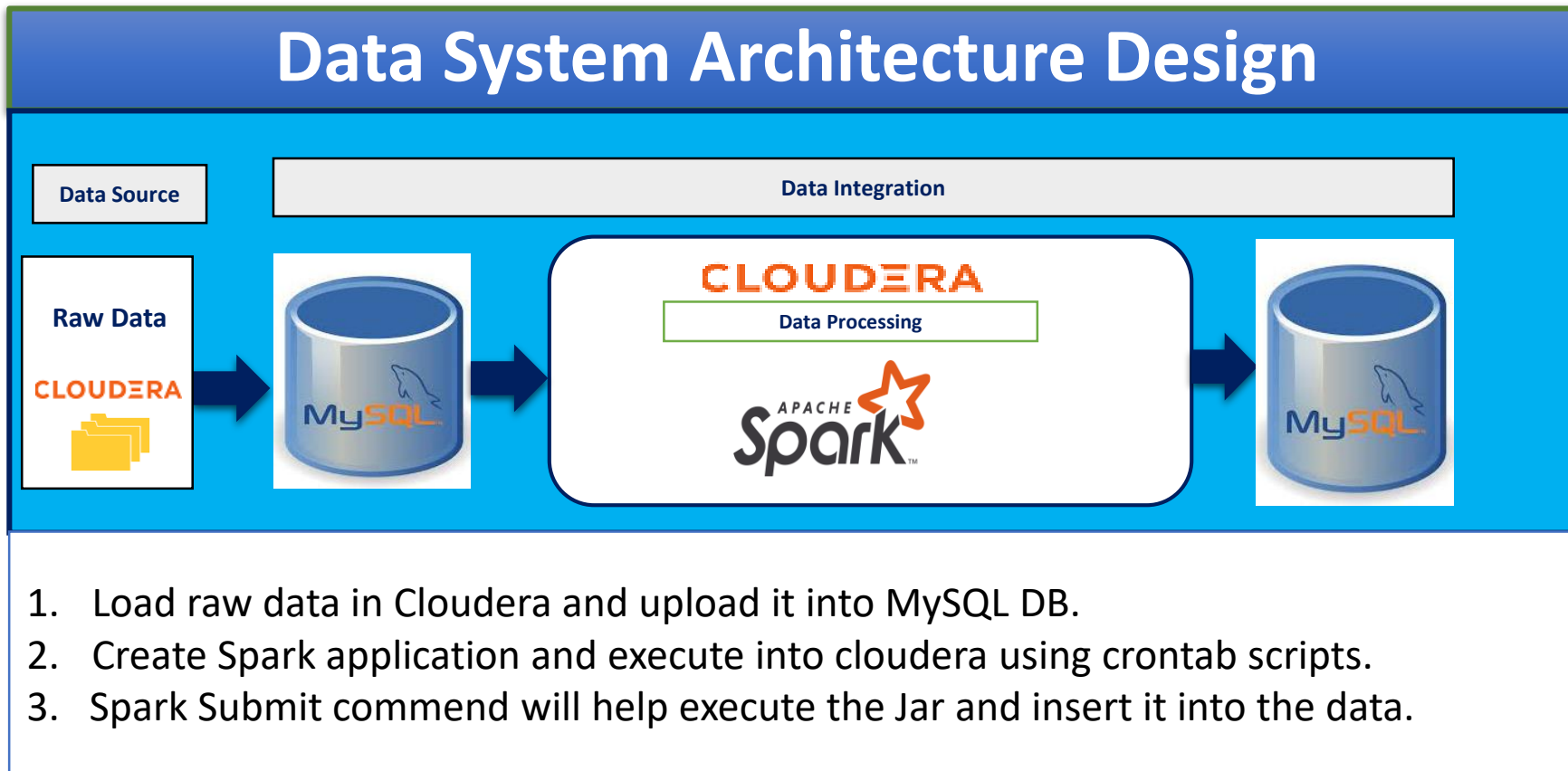
Govetech – Data Pipelines

12-May-2022

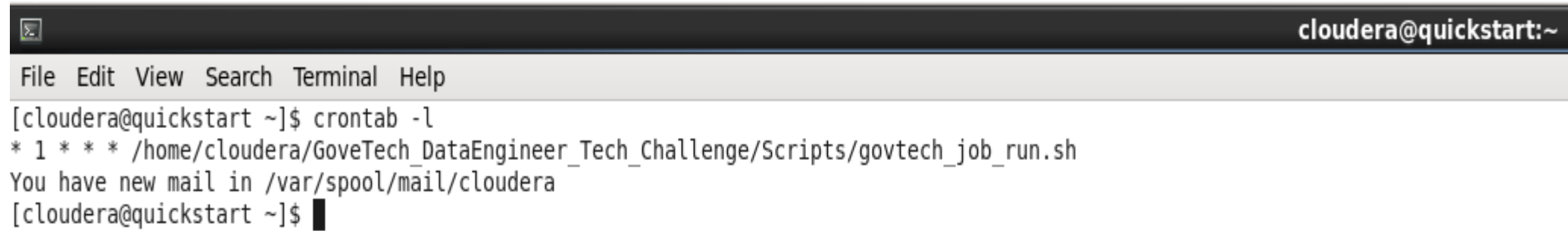
Contents

- Ingestion Files Status Overview
- Crontab Script
- MySQL DB – Output
- Execute Scripts

1. Ingestion Files Status Overview



2. Crontab - Script

A terminal window with a dark title bar showing 'cloudera@quickstart:~'. The menu bar includes 'File', 'Edit', 'View', 'Search', 'Terminal', and 'Help'. The terminal content shows the command '[cloudera@quickstart ~]\$ crontab -l' followed by the output '* 1 * * * /home/cloudera/GoveTech_DataEngineer_Tech_Challenge/Scripts/govtech_job_run.sh' and a notification 'You have new mail in /var/spool/mail/cloudera'. The prompt '[cloudera@quickstart ~]\$' is followed by a black cursor.

```
cloudera@quickstart:~  
File Edit View Search Terminal Help  
[cloudera@quickstart ~]$ crontab -l  
* 1 * * * /home/cloudera/GoveTech_DataEngineer_Tech_Challenge/Scripts/govtech_job_run.sh  
You have new mail in /var/spool/mail/cloudera  
[cloudera@quickstart ~]$
```

3. MySQL DB - Output

```
Applications Places System cloudera@quickstart:~
Window Menu Search Terminal Help
mysql> show tables;
+-----+
| Tables_in_govtech |
+-----+
| govtechuser_price_details |
| govtechuser_price_rawdata |
+-----+
2 rows in set (0.00 sec)

mysql> select * from govtechuser_price_details limit 10;
+-----+-----+-----+-----+
| first_name | last_name | price | above_100 |
+-----+-----+-----+-----+
| William    | Dixon    | 109    | true       |
| Kristen    | Horn     | 263    | true       |
| Kimberly   | Chang    | 187    | true       |
| Mary       | Ball     | 283    | true       |
| Benjamin   | Craig    | 144    | true       |
| Cathy      | Werner   | 62     | false      |
| Brandon    | Bell     | 49     | false      |
| Paul       | Farley   | 12     | false      |
| Sarah      | Mcdaniel | 152    | true       |
| Caroline   | Anderson | 76     | false      |
+-----+-----+-----+-----+
10 rows in set (0.00 sec)

mysql> select * from govtechuser_price_rawdata limit 10;
+-----+-----+
| name      | price |
+-----+-----+
| name      | 0      |
| William Dixon | 109    |
| Kristen Horn  | 263    |
| Kimberly Chang | 187    |
| Mary Ball    | 283    |
| Benjamin Craig | 144    |
| Cathy Werner  | 62     |
| Brandon Bell  | 49     |
| Paul Farley   | 12     |
| Sarah Mcdaniel | 152    |
+-----+-----+
10 rows in set (0.00 sec)

mysql> load data local infile '/home/cloudera/GoveTech_DataEngineer_Tech_Challenge/Inputdata/dataset2.csv' into table govtechuser_price_rawdata fields terminated by',';
```

4. Execute Scripts

```
/home/cloudera/GoveTech_DataEngineer_Tech_Challenge/Scripts/govtech_job_run.sh - cloudera@192.168.121.128 - Editor - WinSCP
```

```
#####
#      Project - Govtech Dataengineer Task      #
#      Created By : Ilango Ramesh Date : 11-May-2022      #
#####
```

```
# Execute spark application
spark-submit --conf "spark.executor.extraClassPath=/home/cloudera/sparkjars/shc-core-1.1.1-1.6-s_2.10.jar" --conf "spark.driver.extraClassPath=/home/cloudera/sparkjars/*"
--class GoveTech_DataEngineer_Tech_Challenge.GoveTech_DataEngineer_Tech_Challenge --master local[*] file:///home/cloudera/GoveTech_DataEngineer_Tech_Challenge/GoveTech_DataEngineer_Tech_Challenge-0.0.1-SNAPSHOT.jar
file:///home/cloudera/Gov_Tech_Rawdata_Processing/Inputdata
```

```
echo " >>>>>>>>Delete user/cloudera - directory output folder - Script <<<<<<<<<"
      hadoop fs -rmr /user/cloudera/GoveTech_DataEngineer_Tech_Challenge/*
```

```
echo " >>>>>>>>>>>>Copy Mysql data to HDFS folder <<<<<<<<<<<<<"
sqoop import --connect jdbc:mysql://localhost:3306/govtech --username root --password cludera --table govtechuser_price_details --m 1 --target-dir /user/cludera/GoveTech_DataEngineer_Tech_Challenge/Outputdata;
```

```
echo " >>>>>>>>>Delete home/cloudera directory output folder - Script <<<<<<<<<<<<"
rm -rf /home/cloudera/GoveTech_DataEngineer_Tech_Challenge/Outputdata/*
```

```
echo " >>>>>>>>>>>>Copy HDFS folder to Local edge node <<<<<<<<<<<<"
hadoop fs -get /user/cloudera/GoveTech_DataEngineer_Tech_Challenge/Outputdata /home/cloudera/GoveTech_DataEngineer_Tech_Challenge/Outputdata
```

```
echo " >>>>>>>>>>>>Delete user/cloudera - directory output folder - Script <<<<<<<<<<<<"
hadoop fs -rmr /user/cloudera/GoveTech_DataEngineer_Tech_Challenge/*
```

```
echo " >>>>>>>Process Completed <<<<<<<<<"
```