Big Data Technologies

Cron

IO Redirection

- output redirection: command > file
- output redirection append: command >> file
- error redirection: command 2> file
- input redirection: command < file
- output on terminal as well as in a file: command | tee file
- error redirection on stdout: 2>&1

CRON

- CRON is process scheduler for Linux.
- It can schedule process execution periodically or at fixed time.
  - Run a task: on 31-Dec-2022 00:00:00
  - Run a task daily: at 12:30 AM
  - Run a task on each Sunday: at 8.00 AM
  - Run a task monthly: on 10th date at 12:00 PM

Write CRON Job

- Install cron on your system. terminal> sudo apt install cron
- Cron job are written in text file using command "crontab -e".
- On first execution, it asks for the editor.
  - Recommended: vim.basic
- Cron job are executed in background by "cron" service/daemon.
  - terminal> sudo systemctl status cron
  - terminal> ps aux | grep "cron"

- Cron expressions: https://crontab.guru/
- Cron example

```
*/2 *   * * *   /usr/bin/date >> /tmp/dates.log
5   0   * * *   ingest.sh
```

- ingest.sh (example)

```
#!/bin/bash

export HADOOP_HOME=/path/of/hadoop
export PATH=$HADOOP_HOME/bin:$HADOOP_HOME/sbin:$PATH

# download weather data from weather station website
/usr/bin/wget https://ncdc/todays -o /tmp/weather.txt

# upload into hive staging table (external table)
hadoop fs -put /tmp/weather.txt /user/nilesh/ncdc_staging
```

Airflow

first_dag.py

- task1 -- Bash Command -- echo "This is First Task."
- task2 -- Bash Command -- echo "This is Second Task."
- task3 -- Bash Command -- echo "This is Third Task."
- Task dependency:
  - task1 --> task2 --> task3
- To execute the DAG
  - Copy it in $AIRFLOW_HOME/dags folder.

- Start the webserver. terminal> airflow webserver
- Start the scheduler. terminal> airflow scheduler
- Open browser: http://localhost:8080/
  - Username: admin, Password: password created while installation.

mysql_dag.py

- task1 -- MySQL command -- CREATE TABLE
- task2 -- MySQL command -- INSERT
- Task dependency:
  - task1 --> task2
- To execute the DAG
  - From airflow web ui, create MySQL connection with appropriate properties.
  - Copy it in $AIRFLOW_HOME/dags folder.
  - Restart scheduler (if dag is not auto-detected).

Security

- Hadoop Security
  - https://youtu.be/Y-f78tjuA6I