September 28, 2022

IS3107 Individual Course Project (Sem1 AY2022/2023)

Data Pipeline Implementation with DBLP Data

Goal

This hands-on project aims to reinforce and integrate the major concepts of data engineering that you have learned from the lectures and tutorials into a basic example of ETL pipeline, based on a dynamic real-world dataset. After completing this project, you will gain a better understanding towards how to leverage (1) open-source data API, (2) DBMS, (3) workflow management tool, and (4) cloud DB services, all together to build a data pipeline.

Background

To better keep track of research progresses, it's a common need for researchers to get regular alerts about the latest publications of other scholars in relative fields. DBLP is a popular computer science bibliography archive that provides open bibliographic information on major academic journals and conference proceedings. Many researchers in the CS community use its reliable open-data services to access and maintain publication profiles, view coauthor information, and obtain links to the electronic editions of publications, etc. DBLP is maintained by a specific team and is updated to include bulks of indexed conf/journal volumes in a daily basis. To automate such kinds of data operations, this project aims to create a data pipeline that regularly imports updated publication data from DBLP to local and cloud databases and fulfil user needs.

Instructions

1. Data Extraction
Use the DBLP person page export API to fetch the publication profile (like the Ooi.xml used in Assignment 1) of each scholar listed in cs_researchers.csv.

2. Data Transformation and Data Storing
2.1 Design and locally deploy a Cassandra database to support the application queries below,

Q1	Given the PID of a researcher, find all the publications of a certain category he/she gets published and signs as the 1st/2nd/3rd… author in a certain year.
Q2	Given the name/PID of a researcher, find the number of times he/she has collaborated with any one of his/her coauthors in a certain year.

Note: You should consider the table design carefully, (i.e., the selection of primary keys) as no "ALLOW FILTERING" is expected for the query execution.

2.2 Deploy an AstraDB (cloud Cassandra) to support,

- A log for volume update of local database, columns include,
    - Timestamp (dd/mm/yy) of the DAG run
    - The number of new publications added
    - The updated total number of unique publications stored in local database
- A log for author-publication updates, columns include,
    - Timestamp (dd/mm/yy) of the DAG run

- Title of the new publication added, which indicates the new publication records updated on DBLP since last check
- "authors" column as the one in Assignment 1, but only include those occurred in the cs_researchers.csv file
- "ee" column (datatype: set collection) as that in Assignment 1, for an easy navigation to available accesses to the new publications

- Storing proxy query outcomes to cloud

Execute the following queries in the local Cassandra DB and store the query outcomes in cloud AstraDB, (Note: You should first create the corresponding tables for these outcomes. This is not a part of the pipeline hence you should only do this for once.)

Q1. Find the total number of conference papers Lihua Xie (PID: 40/2499) has published as the 3rd author during the most recent decade (2012-2022).

Q2. Find the total number of publications Ooi Beng Chin (PID: o/BengChinOoi) has published as the 2nd author during the most recent 5 years (2017-2022).

Q3. How many times in total so far has Lihua Xie collaborated with each of his coauthors?

Q4. Find one of Ooi Beng Chin's coauthors with whom he has collaborated most in year 2020 along with the corresponding times.

3. Data Pipeline

3.1. Check each researcher's DBLP person page in a weekly basis and update the local database accordingly.

- During each DAG run, for each researcher, you should inspect whether he/she has got new publications on DBLP since last check and add those new records to your local Cassandra database and update all the relevant records in your tables accordingly.

3.2. Log the details of update for each DAG run in the tables of cloud AstraDB as described in 2.2.

Submission Requirements

1. Submission content

1) A main report (IS3107Project_YOURNAME.pdf, 3-6 pages, 12pt font, including in-line images, excluding cover page, references, and appendices) to cover the following content:

- Physical Data Model of local Cassandra database
  - You should use the Chebotko notations as introduced in Tutorial 5 to illustrate your Cassandra database design, hence, to **explicitly** specify how you define the primary keys of your tables.
- Your CQL DDL & DML statements
- Snapshots of tables with data for both local and cloud database
- Snapshots of proxy query outcomes stored in your cloud tables
- Describe the rationales that you have considered for your design of the pipeline, such as speed, throughput, reliability, and others mentioned during lectures/tutorials.
- Snapshot of the graph visualization of your pipeline
- Snapshot of the tree view of your pipeline after triggering your DAG
- Run time of each step in your pipeline

2) Airflow DAG file (.py)

2. Submission due at 23:59:59, Monday of Week 14, 14 Nov 2022