SWE 307 BIG DATA PROJECT - 3 DESCRIPTION

Kafka-Spark-Cassandra Data Path

Due date: 13.12.2024 Thursday, in class.

In this project study, you are asked to create a data path that continuously retrieve, process and store data. The architecture of the system is given in Figure 1.

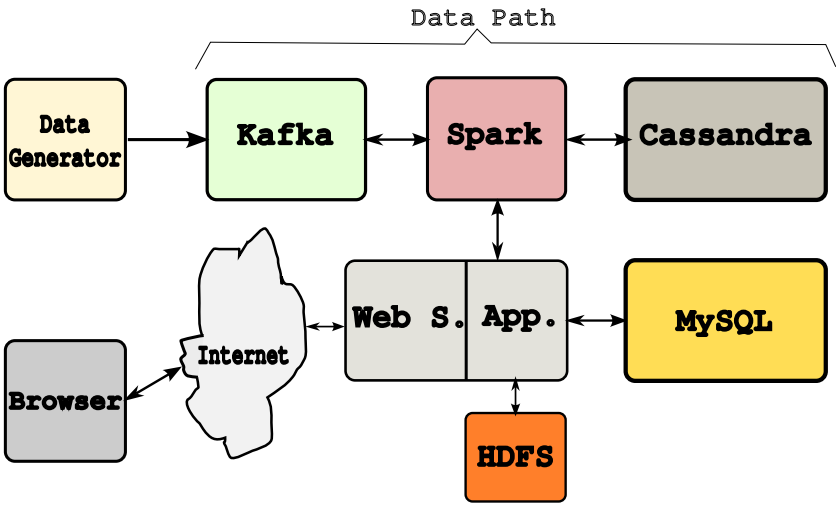

Figure 1. The architecture of the project 3 study.

Here are the descriptions of each block in the figure:

Data generator: This module generates random credit card expenses for the people recorded in MySQL-DB. A new expense record must be generated for every second as in the format given below:

<user_id, date_time, description, type, count, payment>

user_id:	Limited to the emp database.
date_time:	Date and time of the payment.
description, type:	Example “Macaroni, food”, “Jacket, clothe”, “Car, vehicle” etc.
payment:	Amount of money paid, float number, 2 digits precision after the decimal point.

Kafka: This module will gather all expenses generated by data generator. Expense records will be put different queues based on user_ids. Hence, there will be a topic for every user at emp database in MySQL-DB.

Spark: This module gets every data record from Kafka and immediately store them to the Cassandra DB. Spark has also a connectivity with Java-App which creates responses as cumulative expense report for every user queried. For example, if a query sent from the browser for user “Scott” then all information along with toatl expenses must be shown in the browser. In this case, Spark will scan Cassandra-DB, calculate total expense amount and send the answer to the controller.

Cassandra: This is NoSQL database that stores anything sent from Spark module. It can be read from Spark module only for generating instantaneous expense reports.

The lower layer of the system is just a copy of the Project 2. The user information for this project is stored in the "emp.csv" file, you can import them into MySQL-DB. The department information is stored in the "dept.csv" file.

What is required from you is as follows:

1) Hadoop-HDFS cluster must be installed as single node in your computer.

2) A simple Java Spring-Boot application will be developed to perform the following tasks:

a) Personnel and department data will be read from MySQL, calculated expenses will come from Spark (searched, fetched from Cassandra) .

b) Personnel images will be stored to/read from HDFS.

c) Expected web page will show the following information:

<Image of employee, name, mgr. name, salary, commission, department, total_expense>

PS:
1) You are free to use G-Drive or AWS-S3 as file storage instead of HDFS.
2) Example image files and csv files will be provided on Github repository, you can clone/download everything provided.
Link: https://github.com/ozmen54/SWE307-2024.git

Here you are text data as well:
emp.csv
empno,ename,job,mgr,hiredate,sal,comm,deptno,img
7369,SMITH,CLERK,7902,17-DEC-1980,800,,20,smith.jpg
7499,ALLEN,SALESMAN,7698,20-FEB-1981,1600,300,30,allen.jpg
7521,WARD,SALESMAN,7698,22-FEB-1981,1250,500,30,ward.jpg
7566,JONES,MANAGER,7839,2-APR-1981,2975,,20,jones.jpg
7654,MARTIN,SALESMAN,7698,28-SEP-1981,1250,1400,30,martin.jpg
7698,BLAKE,MANAGER,7839,1-MAY-1981,2850,,30,blake.jpg
7782,CLARK,MANAGER,7839,9-JUN-1981,2450,,10,clark.jpg
7788,SCOTT,ANALYST,7566,09-DEC-1982,3000,,20,scott.jpg
7839,KING,PRESIDENT,\000,17-NOV-1981,5000,,10,king.jpg
7844,TURNER,SALESMAN,7698,8-SEP-1981,1500,0,30,turner.jpg
7876,ADAMS,CLERK,7788,12-JAN-1983,1100,,20,adams.jpg
7900,JAMES,CLERK,7698,3-DEC-1981,950,,30,james.jpg
7902,FORD,ANALYST,7566,3-DEC-1981,3000,,20,ford.jpg
7934,MILLER,CLERK,7782,23-JAN-1982,1300,,10,miller.jpg

dept.csv
deptno,dname,loc
10,ACCOUNTING,NEW YORK
20,RESEARCH,DALLAS
30,SALES,CHICAGO
40,OPERATIONS,BOSTON