truecaller

Truecaller

Data Engineer

Take Home Assignment

Position

Senior Data Engineer

Assignment

This assignment should not take more than 5 hours, if it takes more, there may be some sort of misunderstanding. Read the description text carefully and if there still is something unclear, ask your questions via email to: *deepti.pabbisetty@truecaller.com*

Coding Challenge

Let's consider the situation where there are user events coming with millions of records every day. The event contains information about settings changes with this schema:

Column Name	Column Type
id	long
name	string
value	string
timestamp	long

We'd like to run a **daily ETL pipeline** that will read these events, apply some transformation using **Spark** and produce a **partitioned table in Hive**.

The transformation required is to aggregate over the **id** column and merge the **name** and **value** columns to a Map type where the **name** column represents the key and the **value** column represent the value in the Map type.

The value that should be picked for each key in the Map is the one with the highest value from the **timestamp** column.

The output schema should be:

Column Name	Column Type
id	long
settings	Map<string, string>

Example:

User event table:

id	name	value	timestamp
1	notification	true	1546333200
3	refresh	denied	1546334200
2	background	notDetermined	1546333611
3	refresh	4	1546333443
1	notification	false	1546335647
1	background	true	1546333546

Output table:

id	settings
1	{“notification”: “false”, “background”: “true”}
2	{“background”: “notDetermined”}
3	{“refresh”: “denied”}

Note: The Spark job should be as optimized as possible with the assumption that it will work with hundreds of millions of records per run.

Deliverables:

- A runnable Scala Spark project that will apply the transformation.
- Airflow python configuration file describing the ETL pipeline.
    - It should run daily at midnight.
    - It should have a sensor to check for data availability.
    - It should run the Scala Spark job on sensor success.

truecaller

Design Question

Write a short proposal describing the design of a data platform solution for a company with the following assumptions:

- The company's main data source is the mobile application events and backend databases.
- The company's current user base is 20 Million and growing in user volume more than 50% every year.
- Stakeholders using the system should be data analysts and data scientists with minimal engineering experience

Requirements:

1. Data should be served to the stakeholders in a secured and governed way
2. Data Scientists expect a tool to use for accessing the data and build machine learning models
3. Reports should be populated in a BI tool and updated in daily basis
4. Data analysts should be able to do ad hoc queries on the data
5. The project budget is limited and the commercial tools should be minimized as much as possible

Delivery points:

1. A design diagram showing the big picture of the solution.
2. Justify each and every design decision you make, like tools or platform you used, data layouts and modeling, and so on.
3. You are expected to show the flow from collecting the events, ingesting, etl, data modeling, upto the point where data is being consumed by the stakeholders

Thanks in advance for your time and interest in Truecaller.