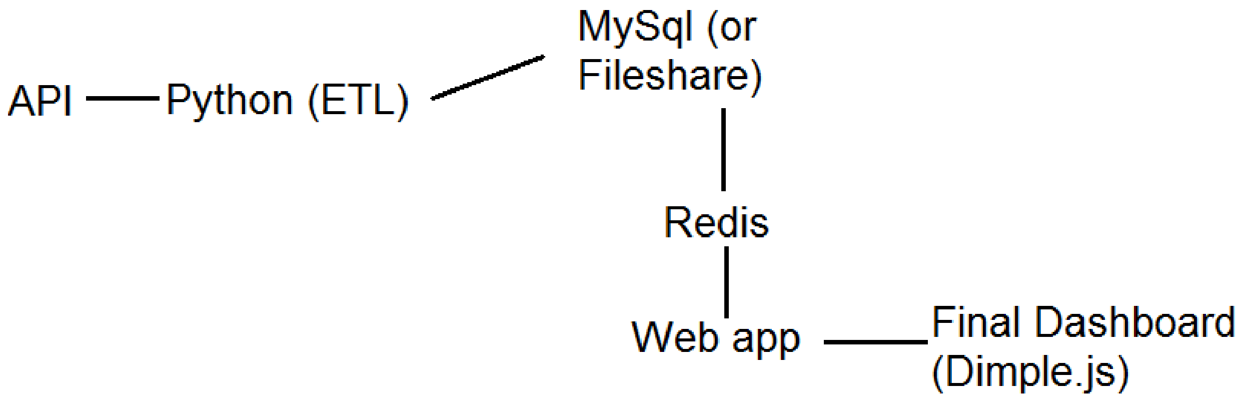
Solution for task 1 can be:



Explanation:

I would go with solution where Python makes ETL (extract data from API, transform – add date to each CSV column- and load data into database). Purpose of Redis is to speed up response of BI dashboard and with it BI app would be very responsive.

Note: I am not sure about logic in which SOAP aggregate data, maybe we always have fully aggregated data not incrementally, so it means that in transformation we should calculate amount of clicks each day.

Example:

01.01.2018 we have

Blogid	views	cliks
1	100	200

Meanwhile someone clicks on our page and tomorrow we have:

Blogid	views	cliks
1	101	201

So, if this is the case we should always save **last snapshot (CSV) of our extraction**, just to be able to calculate in transformation exact number of clicks for some day. // in our case 1 click and 1 view with corresponding date.

Potential problem of this solution:

-What if we have situation where **INPUT RATE > PROCESSING RATE**, for example we have large amount of data and our single Python ETL script is not able to handle it. If it is the case, we could split our Python script on many instances and to have some kind of parallelism in processing data. So, then a couple of Python scripts would work on the same time with chunks (part of CSV file) of data.

Problem with this approach of multiply Python scripts for ETL is scalability. If we want to change something we would need to change it (and test) in each instances of script. It may be really tedious. Also, what if some Python script fail, then it would be hard to concatenate data in final file to import in database, etc.

Generally speaking, due to all this reason, if we really have big data it is maybe the best solution to throw data directly in HDFS of Hadoop, than to make ETL with Map Reduce, and finally to store data in Data Warehouse.

So potential improvement of solution to be considered is:

API – HADOOP – DATAWAREHOUSE – REDIS – FRONTEND OF BI app

Hadoop would be used to store data and to make transformation with Map Reduce and to throw data into data warehouse, then we have in memory database REDIS that would serve our BI app. Advantage of adding Redis is speed.

That is my opinion.