

**Context:**

I own a multi-specialty hospital chain with locations all across the world. My hospital is famous for vaccinations. Patients who come to my hospital (across the globe) will be given a user card with which they can access any of my hospitals in any location.

**Current Status:**

We maintain customers data in Country wise database due to local policies. Now with legal approvals to build centralized data platform, we need our Data engineering team to collate data from individual databases into single source of truth having cleaned standardized data. The Architecture board has chosen Datalake and Databricks approach due to its capabilities with big data solutions. Business wants to generate a simple PowerBI report for top executives summarizing till date vaccination metrics. This report will be published and generated daily for the next 18 months. The 3 metrics mentioned below are required for the phase 1 release.

**Deliverables for assessment:**

Spark (Scala / PySpark) code that does the below

* Data cleansing / exception handling
* Data merging into single source of truth
* Data transformations and aggregations
* Code should have unit testing

**Metrics needed:**

* **Total vaccination count by country and vaccination type**
* **% vaccination in each country (You can assume values for total population)**
* **% vaccination contribution by country (Sum of percentages add up to 100)**

**Expected output format**

* Metric 1: CountryName, VaccinationType, No. of vaccinations
* Metric 2: CountryName, % Vaccinated
* Metric 3: CountryName, % Contribution

*NOTE: End goal is to create data that can be consumed by PowerBI report directly. Candidate is not expected to create the PowerBI report.*

For the assessment, scope is 3 countries. Attached is sample file we will get from each country. Initially you will receive a bulk load file for each country, post that you will receive daily incremental files for each country

**Next Steps**

Once you submit, we will review the deliverables and will reach out to schedule an interview if we like your approach. If we do move forward, we will like to see a demonstration of your working script/program during the interview

We look forward to getting your repo back on [lifi@incubyte.co](https://incubytein-my.sharepoint.com/personal/lifi_incubyte_co/Documents/Microsoft%20Teams%20Chat%20Files/lifi@incubyte.co) and if you have questions at any time, please feel free to send a message or call us on +91-98241-11764

To know more about the company, check out: <https://www.incubyte.co>