

Round 1: Coding Test

Spark Coding Assignment

Prerequisites

This round is heavily based on Apache Spark. So a working installation of Spark is needed. It's upon the candidates to decide whether to code the following exercise in Scala using Spark's [Scaladocs](#) or use Python using [Sphinx](#) (pySpark). The following are the components needed.

- [Apache Spark](#)
- [HDFS](#) // This is optional. HDFS is to read and write the input/output of the program. The candidate can also use local FS.

Below are links for installation of Apache Spark

- [Install Spark on Ubuntu \(PySpark\)](#)
- [Install Apache Spark on Ubuntu 16](#)
- [How to Install Spark on Ubuntu](#)
- [How to install Apache Spark on Windows?](#)

Exercise

Attached below are links to 2 files. Please download these files in your local file system.

- [consumerInternetFile](#)
- [startUpFile](#)

Combine these 2 files in Spark into a single entity and answer the following questions.

NOTE: Answer the below questions using Spark SQL Queries.

- How many startups are there in Pune City?
- How many startups in Pune got their Seed/ Angel Funding?
- What is the total amount raised by startups in Pune City? Hint - use regex_replace to get rid of null
- What are the top 5 Industry_Vertical which has the highest number of startups in India?
- Find the top Investor(by amount) of each year.
- Bonus:
 - Find the top startup(by amount raised) from each city?
 - Which SubVertical had the highest growth(in number of startups) over the years?
 - Which SubVertical had the highest growth(in funding) over the years?

Execution automation instructions

Write the final output of each answer on HDFS/local disk and also print schema of each output. Generate output as mentioned below,

- Create a linux bash script to execute above code
- Once execute it should create output directory as : /tmp/output
- Inside the output directory it should create output of all the answers in different folders.
- Create a README file with steps of compilation, configuration and execution mentioned in it

Java Coding Assignment

Write Java Program to accomplish following,

- Create a Java maven project (details [here](#))
- Read two files mentioned in above
- Instantiate an in memory instance of HSQLdb (details [here](#))
- Connect with in memory HSQLdb with JDBC connection.
- Create two tables inside HSQLdb corresponding to the schema for two mentioned files.
- Load data read into above two tables through Java code (jdbc)
- Execute queries as mentioned in earlier problem statement through jdbc connection
- Output of these queries should be stored in separate output files
- Execution automation expected,
 - Give Java code should take command line arguments as,
 - --output_dir : Output directory path in which to generate output
 - --input_csv : Path of input csv file
 - --input_parquet : Path of parquet file
 - There should be one bash script created on top of above Java CLI
 - Once for execution of bash script, it should create output in output directory /tmp/output_2/ with file names as : 1.csv, 2.csv, 3.csv, 4.csv, 5.csv
 - Create a README file with steps of compilation, configuration and execution mentioned in it

About assignment submission

- Please create your github account if not already created
- Create new git repository inside
- Create two separate directories for each assignment mentioned below
- Commit and push your code to git repository and share across link of same repository as submission



- Mail URL of git repository to ajitr@dataeaze.io, shardul.shinde@dataeaze.io, vishnu.kurup@dataeaze.io, minal.agrawal@iauro.com .