Clairvoyant Interview- 60 mins

***R1 Technical evaluation***

* In the technical R1 - 10 mins - understand the profile
* 45 min- tech evaluation
* 5-10 mins questions from candidate

1.Understanding the profile, the kind of projects done in DA role.

2. Python – Read Write files, Data Structures, Jupiter

Example- Array [0,0,1,0,0,0,0] Maximum length of continuous sub array of 0

3. ETL - Ingestion of data, Transformations, DWH, Data Quality,

4. Communication

5. Data Modeling -Google Meet use case:

6. Questions around analytics

Identifies the basic entities and attributes and ER Diagrams and relationships- 2

Layers of modeling - 3

Best practices in each layer - 4-5

7. SQL - not more than 4 questions. 2 simple, 2 complex - Rating (1-5)

2 simple questions answered – 2 rating

Conceptually knows how to solve complex SQL – 3 rating

4 questions answered - 4-5 rating based on confidence during the answering

Simple:

Joins, Rank functions

Complex:

Scenario based SQL queries (Self joins, Sub queries)

**Evaluation SQL Question (Example)**

**Simple-**

Q1. There are 2 tables with some

Table\_A Table\_B

Id Id

1 1

1 1

1 1

Null

1. Query to get all the records from Table\_A and Table\_B using inner join
2. Output of the above query in terms of number of rows and for different type of joins - left, right, cross, full (whatever the candidate knows)
3. Difference between Join and Inner Join. Left and Left Outer Join. Give us no. of rows.

Q2.

Employee

Emp\_Id Emp\_name Salary Dep\_Id

1 Ravi 100 1

2 Seema 200 2

3 Rahul 100 1

4 Shreya 200 1

5 Ramesh 300 2

6 Raj 300 2

Department

Dept\_Id Dept\_Name

1 IT

2 HR

1. Write O/P with 3 more columns as Rank(), Dense\_Rank() and Row\_Number()
2. Query to find all the employees whose salary is between 200 and 300
3. Query to find the name of the employees that begin with ‘R’
4. Write a query to retrieve departments who have less than 2 employees working in it.

**Medium**

Q1

Employee

Emp\_Id First\_Name Last\_Name Gender Salary Dept\_Id

1 Ravi Goyal M 100 1

2 Rahul Bansal M 200 2

3 Mohan Kumar M 150 1

4 Shyam Mohan M 200 2

5 Lata Mohan F 150 1

6 Reema Singh F 100 2

Department

Dept\_Id Dept\_Name

1 IT

2 HR

3 Sales

1. Query to get all the records of the employees where salary is maximum in each of the departments.
   1. First\_Name Last\_Name Gender Salary Dept\_Name
2. Query to get all the employees records with the 2nd highest salary.
   1. First\_Name Last\_Name Gender Salary Dept\_Name

**Complex**

Q1

Table Name-Team

Id Team\_name

1 India

2 Pakistan

3 SriLanka

4 Africa

5 Australia

O/P (All combinations)

India - Pakistan

India - SriLanka

India - Africa

India - Australia

Pakistan - SriLanka

Pakistan - Africa

Pakistan - Australia

SriLanka - Africa

SriLanka - Australia

Africa - Australia

i) Query to get all the above combinations.

Q2.

Input

Table - Student

Id Student

1 Mark

2 David

3 John

4 Albert

Table - Friends

Id Friend\_Id

1 2

2 3

3 4

4 1

Table - Package

Id Salary

1 18

2 12

3 13

4 15

Output

Student\_Name

David

John

Albert

Query to find the students whose salary is less than their best friend’s salary

Q3.

Input

Table- Fund

Fund\_Name Score Category

ICICI 80 A

SBI 70 A

Reliance 60 A

Axis 50 A

IDFC 40 A

ICICI 40 B

SBI 50 B

Axis 75 B

Nippon 40 B

Output - Write a dynamic query that will give the below records with the same number of records as that of input.

Fund\_Name Category Avg score (in the same category leaving its own score)

ICICI A (70 + 60 + 50 + 40) / 4

SBI A (80 + 60 + 50 + 40) / 4

.

.

Nippon B (40+50+75)/3

8. Banking - Domain knowledge

– **Amol’s added questions:**

My questions:

1. Can you tell me a situation where:
   1. Where data available was not enough.
   2. How do you map the data to answer business questions at hand
   3. Any KPIs you developed and how do you identify this as a KPI. What are the supporting parameters, you look at KPI when you are trying to depict the state of the scope or company.
   4. When you knew that you have not worked on something before. How did you communicate this to the stakeholders and how did you approach the problem
   5. When you missed a deadline
   6. When you lead a team in absence of your supervisor/manager
2. Case study questions:
   1. You want to launch a new product or service, how will see if your product or service be a good market fit
   2. You see a sudden change in some product or service, how do you check for the cause
      1. Either sales in an ecommerce store
      2. Signups for your product or service
      3. Users logged in or usage has dropped
3. Behavioral questions:
   1. You see a potential for improvement, how do you approach it
   2. Your colleague is not pulling his weight, how do you work it out.
      1. Either the skills are not enough
      2. Or he is negligent
   3. You have not been given credit for your ideas or your work, but your boss takes it. How do you respond?
   4. Strengths & weaknesses
   5. What do you think is the future of Data Analytics? How do you see yourself in this position?
   6. How do you work best: in a team or individual
   7. What did you do when your stakeholders are non-technical & you are presenting a very technical project to them
4. SQL questions:
   * 1. Present the unique list of origin & destination with the max/min dist:

|  |  |  |
| --- | --- | --- |
| origin | destination | distance |
| Mumbai | delhi | 300 |
| Delhi | Mumbai | 290 |
| Delhi | Pune | 320 |
| Pune | Chennai | 450 |

* + 1. Taking the same question to a more complex level: Present a list with min & max distances for unique origin & destinations
  1. Go the other way around: Give a business problem & ask them to create the data & then a query for that:
     1. For ex. We are looking for customers who have their 2nd transaction more than 1000 rs

1. Could you describe your typical day at your current job
   1. Daily activities
   2. People working with/stakeholders

HSBC Interview Preparation

General Expectation in HSBC interviews:

1. Good communication skills
2. Experience in working with stakeholders, leadership if possible
3. Data pipelines/modeling & automation experience
4. Finance background

Things to prepare for the interview:

1. Crisp introduction
   1. Mention your education/place
   2. Various tech that you have worked on
   3. Some domains that you have worked on
   4. Some analytical projects like fraud detection, defaulter analysis etc
2. 2 projects as introduction:
   1. One could be in detail and closely related to finance/banking (Look at the fig below for this analytics projects)
      1. Mention the business question/objective
      2. How you collected data
      3. Exploratory data analysis/ Pre analysis
      4. Communication with stakeholders
      5. End reports/recommendations
      6. Automation
   2. Other one is more for a backup, just in case you run out of words or finish too early
      1. Same as above may not be necessary, but prepare a 1/ 2 liners summary and if needed you can go in depth, so both are fine

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Phase 1:** Collect data from raw layer | **Phase 2:**  Data cleansing & structuring as much as possible. Also, data modeling in some cases | **Phase 3:**  Integrate into 1 table/ schema/ data model | **Phase 4:**  Data added to appropriate assets/ departments like transactions, customers etc.  MDA | **Phase 5:**  Data converged on different silos like frauds, customer surveys etc | **Phase 6:**  Data available for End users to download/ see |

|  |  |  |  |
| --- | --- | --- | --- |
| Metadata management | Etl pipelines management | User access | Data cleansing standards |
| Keep account of all data available |  |  |  |

1. Behavioral questions:
   1. Strengths & weaknesses
   2. Tell me about a time when you:
      1. Did not finish on time
      2. Helped a coworker
      3. Learned something new on the go
      4. Have multiple deadlines
      5. Etc. project related questions
   3. Leadership experience perhaps for Sr data analyst or higher
2. Technical questions:
   1. SQL/Hadoop/Hive/Spark
      1. Any experience with big data, how big was data
      2. Have you used hadoop anytime? Or spark
         1. Usually if you know SQL you can mention you have worked on hadoop via hive which is SQL based
      3. Data modeling:
         1. What is a conceptual/logical/physical data model
         2. Build a data model for ‘X’ scenario
            1. We have patient/insurance/claims
            2. We have user/account/services etc.
         3. How will you reduce redundancy in data models
            1. Answer is normalization. Simple words if a column has repeated information, you can make a lookup table with id and link the ids here
   2. Project based follow up questions could be, one liners are more than enough here I can say.
      1. How did you follow up with stakeholders and decide what the end deliverable would be
      2. How many file formats the data was distributed across
      3. How did you clean your data or validate it
      4. For ML models:
         1. Where was it deployed and how it was used
         2. How was it updated and used?
         3. How did your users use it.
      5. Data pipelines:
         1. ETL tools used?
         2. Python & cron jobs (learn what is cron if possible, 10 mins and its very simple)
         3. Start files/tables in db to end in transformations
            1. Follow the exact transform load approach. This will give a structure to the answer as well
         4. How often things were refreshed
         5. How big was the data?

1. What you do to upgrade your skill set?

2. Why we use Pyspark not pandas?

3. How will you explain non technical person about data?

4. Why we use Group By?

5. To get all data from set 1 and some data from set 2 which join you will use and which set you will put where.

6.Why we are using Hadoop over GCP, AWS

7. What is EDA?

8. When we stop to using Excel?

9. What will you do when you get unexpected result ?

10. Suppose we have user form which have age input how do you validate it ?

11. You got your data of banking with customer id . What you will do with that id ?

12. How you will deal with null records?

1. What are advantages of hadoop hdfs over gcp, aws storage services?

2. How will you explain some technical process/issue to non technical person?

3. How will you make python strongly typed?

4. How did you deal with some unexpected results in your analysis process?

5. How will you deal with null values in the data?

6. What is EDA. How and what does it help with? Have you done EDA, explain with some example

7. Lets say some customer is in disagreement with replication factor that it leads to more storage requirements to store same data, how will you convince them regarding the significance of replication factor?

8. What technologies you have used in analytics?

9. How do you keep learning new things and acquiring new skills?

10. What size of data have you worked with?

1. Describe about the past data analytics projects you have done ?  
2. Data Modeling concepts like ER diagram and problem statements around it ?  
3. What are your strengths and weakness ?  
4. If someone calls your manager, what are the things he/she would tell about you could improve ?  
5. What motivates you to continue in data world ?

Have you worked on Jira and Confluence  
Given the Big data how you will handle the Data cleansing.  
What is difference between Data cleansing and validation.  
How you would deal with the data having missing values.  
What are the data validations you have implemented in the your project  
Have you worked on Hadoop and GCP  
Have you worked on Data Modeling and Data Profiling  
 How you have worked following Agile methodology  
 How you will ensure Data Quality and Data Integrity is maintained in the data pipeline process  
 Are you comfortable working with UK shift timing  
 Write a SQL query to from the details mentioned showing the total balance of Customer name in descending order

# Tips for a smooth interaction:

1. Maintain Eye contact with the camera
2. Try to have a format to your answered questions.
   1. You can follow the STAR format from Amazon or you can have something of your own
3. Maintain good lighting conditions
4. Try not to have very long answers
5. In case there are any weather warnings, try to give out a warning beforehand
6. Try to have as many mocks as possible with different people.
7. Rest of the usual ones apply like being on time and all :)
8. Please prepare to be able to present yourself in the below format. Here is an outline. See that you create a storyline like this.

Objective is to make an impression that you are the right Data Analyst they are looking for and you know the high level stuff of how data platforms are architected.

**Introduction should cover most of the aspects. So try to come up with a good story telling. It should feel like you are walking them through your journey across your career and one project in particular.**

* Start with Education , Location, Years, Companies, Domains, Technologies -> Not more than 2 mins
* One project Explanation: - 5 mins.
* Problem Statement
* Solution Outline (Data Platform Architecture End to end)
* Sources (Format of data - structure, unstructured, internal and external)
* Raw Layer
* Enriched Layer
* Reporting Layer
* Visualisation
* Your role: - 3 mins
* Data Analysis (techniques and technologies)
* Data Modelling and ETL (if possible) -> high level concepts
* Data Quality Checks -> high level concepts
* Talk about technologies used (esp Python, Pyspark, Hadoop, SQL) - 1 min
* Show your readiness: - 1 mins
* That you are ready to learn GCP, Data Science if possible etc..
* That you are ready to learn Banking

**Few tips and questions to be ready with to be able to answer the questions from the interviewer.**

* Explain the use cases before the client asks. Talk about the domain data assets that were used for the analysis.
* What are some of the challenges while doing the analysis.?
* Avoiding duplicates in dimensions data, fact data .
* What is Data Pipeline?
* Extract from source to target, transformation (data model, cleaning, refining, business logic and making it reusable data sets) and load the data into target.
* Building the insights from the data that can be used by monitoring and decision making teams.
* Notifications ->
* How will the operations teams be notified of failures?
* How will the business users be notified when KPIs are not met?
* What is Data Modeling?
* What is a Conceptual Data Model?
* What is a Logical Data Model?
* What is a Physical Data Model?
* What is the diff between Logical to Physical Mapping?
* What will you do when you have new columns added??
* Check the data dictionary
* Work with source systems owners.
* Check for all the layers of the Data Platform where the columns will have to be added.
* Discuss with all the stakeholders using that table and get their agreement
* Work with the Data Modelers to ensure the changes are good to go from a data modelling standpoint.
* Work with the engineering teams to get the implementation done.
* Load the historic data for those columns if required.
* What are the Data Quality measures?
* Consistency
* Completeness
* Timeliness
* Business Checks that are specific countries/jurisdictions etc.
* What is your experience on Agile??

**STEPS FOR DA-**

As Data Analyst role:

* I usually spent time understanding business requirements and deriving KPI for my analysis. (from Stakeholders).
* Identify the data sources involved and their schema, relationships
* Plan to bring data onboarding to Hadoop platform (if they are outside of the hadoop)
* Create high level documents mentioning a complete plan of analysis including use case, expectations, data sources and relationship and insights that I am planning to get out of it.
* Once I get access to data in impala/ Hive, I start exploring that using HUE interface, verifying my expectations around data like type, boundaries sort of data profiling
* Because our cluster is shared and we have to be careful about resources, I used to download sample data in my local system and to analyze on my machine using pyspark. This also helps me to not be dependent on the data engineering team for data ingestions.
* Once I format data as per my needs like date extractions, new column derivations, extract required information from fields, I used to verify the script end to end and make sure it does what is expected. (DATA DOCUMENTATION IN CONFLUENCE)
* I also helped/ worked in/on scripting Data validation expectations using **PyDeequ** where engineering team later automate it during data onboarding
* I push this code to github or upload my code to cluster jupyter hub, change the spark redirection to spark cluster and local file inputs to impala tables and re-run my analysis, finally save that results to separate workspace databases where I have write access to.
* On cluster run, I used to validate data and report generate using pydeeque to make sure data is following thought expectations
* Once I am confident about data and my expectations, I used to connect tableau to impala database and used to generate insightful dashboards for my use cases

## Good Luck to all!