Big Data Interview Questions

1. How would you go about a Data Analytics Project?

A candidate must know the five key steps to an analytics project:

- **Data Exploration:** Identify the core business problem. Identify the potential data dimensions that are impactful. Set up databases (often using technologies such as Hadoop) to collect 'Big data' from all such sources.
- **Data Preparation:** Using queries and tools, begin to extract the data and look for outliers. Drop them from the primary data set as they represent abnormalities which are difficult to model/predict.
- **Data Modelling:** Next start preparing a data model. Tools such as SPSS, R, SAS, or even MS Excel may be used. Various regression models and statistical techniques need to be explored to come up with a plausible model.
- **Validation:** Once a rough model is in place, use some of the later data to test it. Modifications may be made accordingly.
- **Implementation & Tracking:** Finally, the validated model needs to be deployed through processes & systems. Ongoing monitoring is required to check for deviations; so that further refinements may be made.

2. What kind of projects have you worked on?

Typically, a candidate is expected to know the entire life cycle of a data analytics project. However, more than the implementation, the focus should be on tangible insights that were extracted post implementation. Some examples are:

- The sales data of an organization – Perhaps there was a problem regarding under achievement of targets during certain 'lean periods.' How did you pin the sales outcome to influencing factors? What were the steps you took to 'deflate' the data for seasonal variations? Perhaps you then setup an environment to feed the 'clean' past data and simulate various models. In the end, once you can predict/pinpoint problem factors, what were the business recommendations that were made to the management?
- Another one could be considering production data. Was there a way to predict defects in the production process? Delve deep into how the production data of an organization was collated and 'massaged' to conduct modeling. At the end of the project perhaps some tolerance limits were identified for the process. At

any point, if the production process were to breach the limits, the likelihood of defects would rise – thereby raising a management alarm.

The objective is to think of innovative applications of data analytics and talk of the process undertaken; from raw data processing to meaningful business insights.

3. What are some problems you are likely to face?

To judge how hands-on you are with data and technologies, the interviewer may want to know some of the practical problems you are likely to face and how you solved them. Below is a ready reckoner:

- **Common Misspelling:** In a 'Big Data' environment there is likely to be common variations of the same spelling. The solution is to identify a baseline and replace all instances with the same.
- **Duplicate Entries:** Often a common problem with master data is 'multiple instances of the same truth.' To solve this, merge and consolidate all the entries that are logically the same.
- **Missing Values:** This is easy to deal with in 'Big Data.' Since the volume of records/ data points is very high, all missing values may be safely dropped without affecting the overall outcome.

4. What are your Technical Competencies?

Do your homework well. Read the organization profile carefully. Try to map your skill sets with those technologies that the company uses in terms of big data analytics. Consider speaking about these particular tools/technologies.

The interviewer will always ask you regarding your proficiency with big data and technologies. At a logical level, break down the question into a few dimensions:

- From the programming angle, Hadoop and MapReduce are well-known frameworks generated by Apache for processing large data set for application in a distributed computing environment. Standard SQL queries are used to interact with the data.
- For the actual modeling of the data, statistical packages like R and SPSS are safe bets.

- Finally, for visualization, Tableau and variants like Apache Zeppelin are industry highlights.

5. Your end user has difficulty understanding how the model works and the insights it can reveal. What do you do?

Most big data analysts come from diverse backgrounds belonging to statistics, engineering, computer science, and business. It will take strong soft skills to integrate all of them onto a common page. As a candidate, you should be able to exhibit strong people and communications skills. An empathetic understanding of problems and acumen to grasp a business issue will be strongly appreciated. For a non-technical person, the recommended solution is not to have the Analyst delve into the workings of the model, instead focus on the outputs and how they help in taking better business decisions.