**Interview Transcript: Maneesh Gorantala**

Level: L1 Discussion

Role: Data Engineer

Project: Google US

**Interviewer:** Hi Maneesh, can you start with a quick introduction about yourself?

**Candidate:** Sure. I'm Maneesh Gorantala. I have around 5 years of experience in IT, majorly working on Big Data and Data Engineering projects. I’ve worked with Hadoop, Hive, Spark, and have experience in cloud platforms like GCP and AWS. Most recently, I’ve been working on real-time and batch data processing using PySpark and scheduling pipelines using Airflow.

**Interviewer:** Great. Can you walk me through your current project?

**Candidate:** Yes. I'm currently working on a GCP-based data lake project. We receive raw data files in GCS, then we clean, transform, and enrich the data using PySpark. Finally, we write the curated data into BigQuery for reporting. We use Airflow to schedule the jobs and Stackdriver for monitoring.

**Interviewer:** What kind of transformations do you perform in PySpark?

**Candidate:** We perform column-wise transformations, apply business rules, handle null values, and sometimes perform joins across datasets. In some cases, we use window functions, and we also partition data based on business logic before writing to BigQuery.

**Interviewer:** How do you manage pipeline failures or job retries?

**Candidate:** In Airflow, we configure retries and alert mechanisms. For critical pipelines, we have SLAs defined. Stackdriver sends alerts on failures. We also have logic to identify partial loads and reprocess only the failed partitions.

**Interviewer:** Do you have any experience with streaming data?

**Candidate:** Yes, in one of my previous projects, we used Pub/Sub and Dataflow for real-time processing. We ingested clickstream data and pushed it to BigQuery in near real-time for dashboarding.

**Interviewer:** What challenges have you faced in data processing and how did you solve them?

**Candidate:** One challenge was data skew during joins in Spark jobs. To handle this, we used salting techniques and broadcast joins where applicable. Another challenge was schema mismatches between environments. For that, we implemented schema validation and automated tests as part of our CI/CD pipeline.

**Interviewer:** How comfortable are you with writing SQL queries?

**Candidate:** Very comfortable. I write SQL queries daily for validation, reporting, and debugging. I’m well-versed in window functions, CTEs, subqueries, and optimization techniques like partitioning and clustering in BigQuery.

**Interviewer:** How do you manage schema changes in BigQuery?

**Candidate:** We handle schema evolution through code. For non-breaking changes like adding new nullable fields, we use ALTER TABLE. For breaking changes, we create new versioned tables. We ensure schema compatibility before deploying using automated scripts.

**Interviewer:** Are you involved in DevOps or CI/CD?

**Candidate:** Yes. We use Git for version control, and Cloud Build for CI/CD. All our DAGs and PySpark scripts are deployed through pipelines. We also perform unit testing and deploy across dev, QA, and prod environments.

**Interviewer:** Do you interact with onshore teams or clients?

**Candidate:** Yes, regularly. I attend daily standups, sprint planning meetings, and demos with onshore stakeholders. I also communicate via email or calls for clarifications and status updates.

**Interviewer:** What is your notice period and salary expectation?

**Candidate:** My notice period is 15 days. Regarding salary, I’m looking for a competitive package as per company norms, and I’m open to discussion.