VoiceMod theoretical Answers

1)

Pipeline :

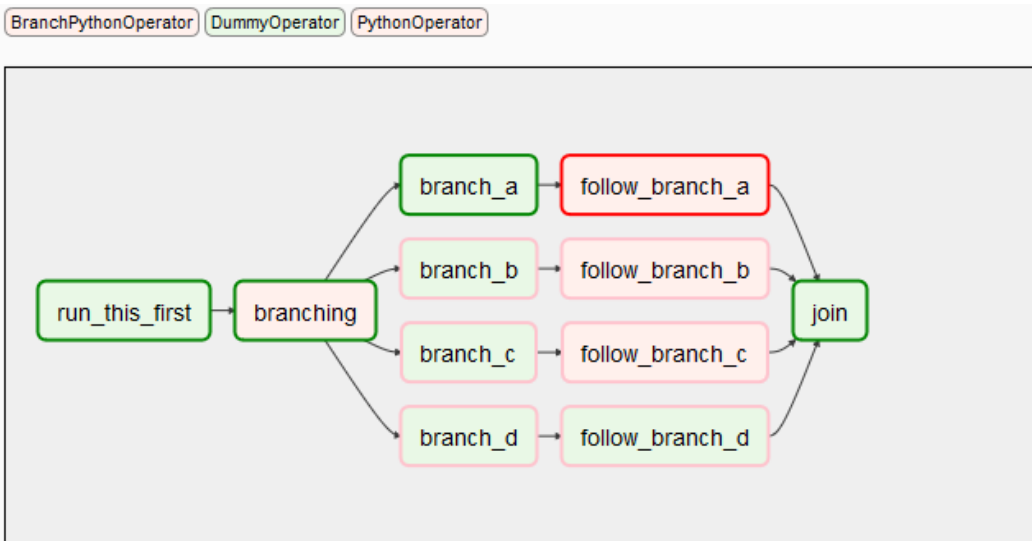
1 - Send periodically buckets of data to S3 (Amazon Web Services). Data should come from backend database from web and desktop application;

2 – From S3, after a period that makes sense (ex: daily), do the ETL with SQL queries and organize the data in first layer schema.

3 – Create different schema like raw, public, analyzes. Each with different level of aggregation and analysts only ad-hoc query in the last one. For this, write sql aggregation query to create their KPI's metrics to build dashboards.

Orchestration: One of the best tools to orchestrate pipelines nowadays is Apache Airflow. I would use it rather than cronServer. In Airflow we could organize pipelines by metrics. For example : User_table since received till the point it is in the last schema for analysts.

(only for example purpose:)



2) Let me assume that marketing team want to see their data. If we send the redshift data they can see ALL data what isn't good. In Redshift we can have nodes in a cluster. We can put one node for marketing and they could see it for example in metabase. The only change in the previous pipeline orchestration would be to hava a DAG per department. This would be

more a work of separate node and understand the needs inside the clusters. The DAG is what is shown in the image. It means Directed Acyclic Graph.

3)Yes. With not structure data the pipeline would change a bit but, Airflow can orchestrate Data Lake too.