Project Abstract, What Mic Should I Use?

Hello and thanks for checking in! Having recently completed IBM’s Data Engineering Professional course, I’m eager to apply what I’ve learned to a personal project involving a critical tool of my past trade as an audio engineer: microphones! I’ve only just begun, so there’s not much to see as of yet, but I’ll be adding to this repository regularly and encourage you to check in occasionally for updates. Because I’ll be working in an environment that is local to my machine, I’ll be using screenshots and pdf’s of reports generated with Power BI to document progress.

My goal is to develop programs, a database/data warehouse, and automated pipelines to generate source data for an analytics dashboard and Apache Spark machine learning algorithm to answer the question, “What mic should I use?” The intended specific context is the amplification/reinforcement of live music. Because the means is more important to me than the end (at least for now), I make no promises that the results will actually provide even remotely scientific evidence of what mic should be used for what instrument in any given situation! Most of the data will be at least partially, randomly generated for lack of an extant source of data available as fuel for my idea. However, I hope to involve the community of audio engineers I’ve gotten to know over the years to ultimately provide a resource of actual value.

To begin, I’ll outline my plan for the basic components of the project:

1. An ERD created with Postgres utilizing a star schema (possibly snowflake schema?), along with a corresponding .sql script.
2. A MySQL staging database/data warehouse with the above Postgres .sql script as schema source.
3. A Python program to randomly generate data for analysis on a daily basis. This program will run via Apache Airflow, i.e. a ‘DAG’, or directed acyclic graph and update the staging database with the new records.
4. A DB2 production database/data warehouse. This database will update daily from the staging database through a Python script via Apace Airflow (i.e., a separate, 2nd DAG).
5. An analytics dashboard connected to the production database. I haven’t chosen the platform just yet: IBM’s Cognos or Microsoft’s Power BI.
6. An Apache Spark job, likely triggered daily by a 3rd Airflow DAG(?), that will utilize machine learning to make a classification prediction regarding what mic should be used for what sound source in what situational context. For the moment, I plan to code this with Python in a Jupyter Notebook running on IBM’s Watson Studio in a serverless context… the job will spin up, run, and shut down once per day.

* David