Review Document
AWS Automated Data Pipeline

Ashutosh
Brhat BG

1st Version with EMR

- **Schema Management:** While the schema is hardcoded within the PySpark job.
- **Dynamic Parameters:** The Lambda function passes bucket locations dynamically to the PySpark job.

2nd Version with EMR

- **Schema Validation in Data Integration:** Extracting the schema from an existing CSV file and ensuring it aligns with the schema of the target CSV file for accurate data merging post-processing.
- **Dynamic Parameters:** Using boto3 sdk in pyspark to dynamically get the csv file names from S3 buckets.

3rd Version with EMR

- **Schema Inference in Data Processing:** The project's data processing strategy allows for schema inference, which can be derived from the new incoming data or the existing datasets. This approach ensures flexibility and adaptability in handling data with potentially evolving structures.

1st Version Glue

- **Crawler:** Used for to infer the schema of the table.
- Crawler runs once before glue job definition.
- Crawler Dynamically reads csv file name in S3 and creates table in glue database.

2nd Version with Glue

- **Schema Management:** While the schema is hardcoded within the glue job.

Learning Outcome

- Team Collaboration
- Cloud Infrastructure Design
- Event-Driven Architecture
- Scalability and Resource Management
- Automation and Scripting
- Monitoring and Logging
- Unit Testing

Thank You!