DSC680 – 8.2 – Milestone 3

Binay P Jena

- Scenario
  - First hire on a data team
  - data lead guidance missing
  - Lack of expertise with infra related aspects

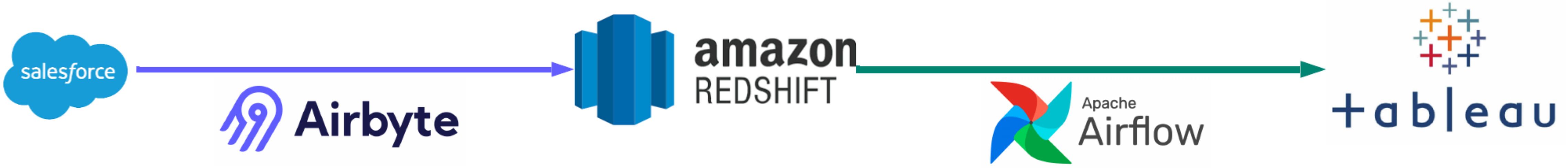
- Problem Statement
  - Where do I start? How to automate DataOps stack?
  - How much would it cost? Is this frugal? Will it scale? Will there be more $ with scaling?

- Assumptions/ Prerequisites
  - Coding language – python
  - Data manipulation – SQL
  - Access to cloud provider (AWS, Azure, etc)
  - Analytic datastore (Redshift, Snowflake, etc)

Solution / Setup Approach

- So many choices, so many options – even by just searching online, one is spoilt for options … an indicative list being
  - Open-source
  - one cloud
  - Hadoop, Hive, Snowflake, Databricks, Redshift, Firebolt
  - Python, Scala, Java, R
  - Notebooks, Scripts, Auto run, Env Variables
  - CI/CD – Jenkins, Jfrog, CircleCI, Gitlab
  - Modern Data Architecture
- This project provides a baseline data stack to get things operational frugally

Data Flow Overview – Components



For detailed workings -

- We use Salesforce Opportunity object to produce a weekly trends of leads and its corresponding revenue using the stack setup.

- Please refer the workings document, and codebase https://github.com/bpjena/ms-dsc/tree/master/dsc680_project for the detailed code workings and instructions to follow.