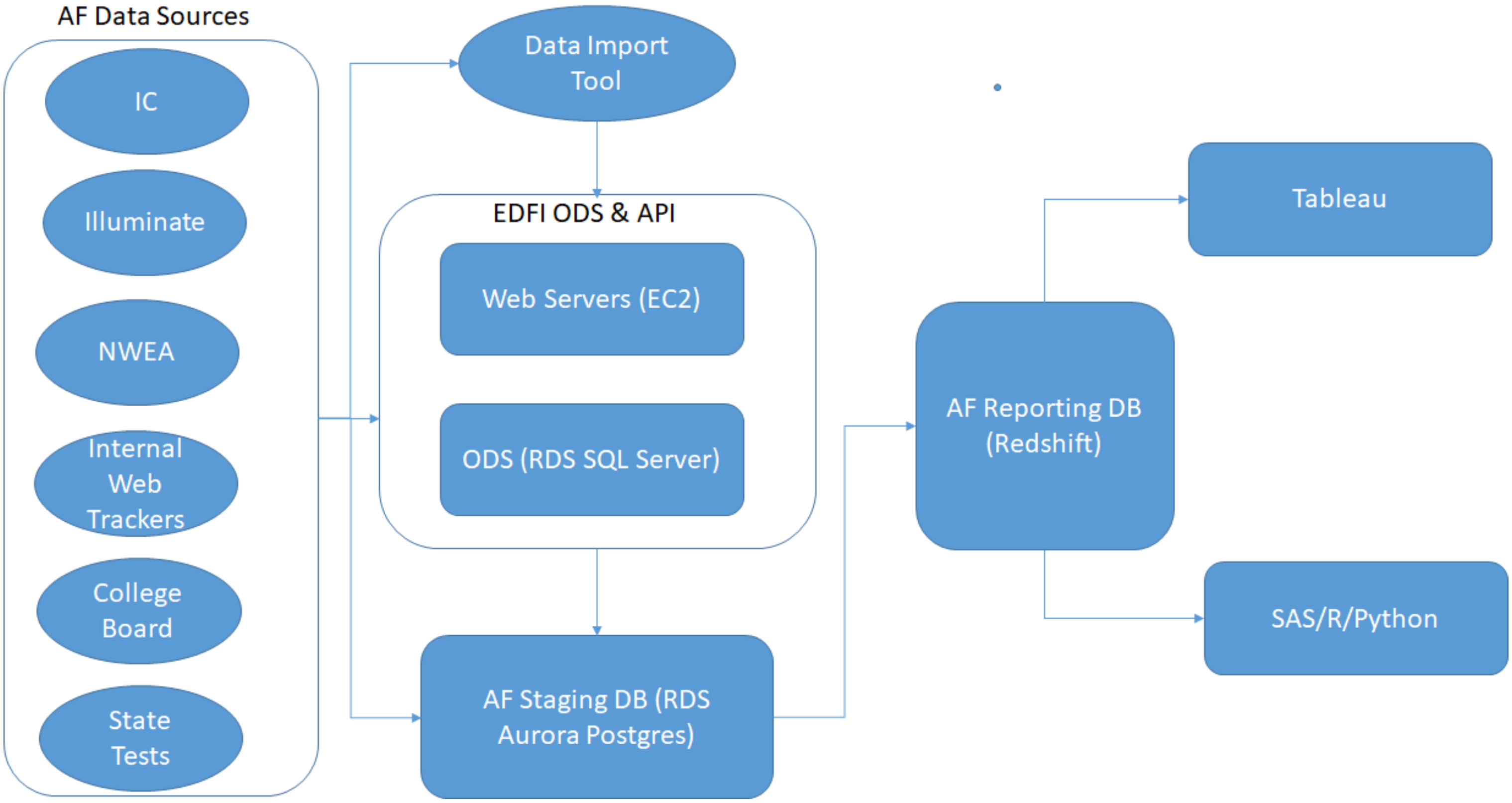
Data Architecture Overview

Summary

AF's data infrastructure is hosted on Amazon Web Services (AWS). This infrastructure includes an implementation Ed-Fi ODS version 3.3.

ETL Framework

The following diagram represents the overall ETL framework deployed in AWS. In this framework, source data is first loaded into the Ed-Fi ODS and then transferred to the staging database. Additional datasets including custom datasets and ad-hoc data trackers are loaded directly into the staging database. Finally, this data is transferred to a central reporting database that powers data visualizations and data analysis.
Source data is loaded into the Ed-Fi ODS using either vendor supported native integration or the data import tool. Additionally, this framework also includes an implementation of the assessment bridge API to translate native Ed-Fi v2.5 integration to Ed-Fi v3.



Overview of AWS services

The ETL framework outlined above uses the following services in AWS.

RDS	The Ed-Fi ODS v3.3 is deployed using the RDS SQL server engine and the staging database is deployed using Amazon Aurora with Postgres compatibility
Amazon Redshift	The reporting database is deployed using this service

EC2	The Ed-Fi API v3.3, Ed-Fi admin app and the assessment bridge API are deployed using EC2
Lambda	All ETL data pipelines are deployed using Lambda. These pipelines are written using Python version 3.6.
Step Functions	The entire ETL workflow outlined above in enabled using Step Functions
S3	This service is primarily used to store raw data files in csv format.

In addition to the primary services listed above, this architecture uses some supplementary services in AWS such as Simple Notification service, Secrets Manager and some networking services. The Ed-Fi ODS and API are deployed using the AWS deployment template for version 3 in its own virtual private cloud (vpc). All other services (non-Ed-Fi) are hosted in a separate virtual private cloud (vpc) with vpc peering enabled between the two instances.

Database Overview

- Staging Database – Contains the Ed-Fi v 3.3 schema and additional schemas to store custom AF data.
- Redshift Database – Contains a separate schema for each dataset.

Git Repository

The starter code in Python for the data pipelines in this framework and the reporting database can be found here: https://github.com/achievementfirst/af-edfi-shared

- The data pipeline code to transfer student and enrollment data from the ODS to the staging & reporting db can be found in the following directories:
  - ODS – Staging – IC
  - Staging – Reporting – IC
- The data pipeline code to transfer assessment data from the ODS to the staging & reporting db can be found in the following directories:
  - ODS – Staging – Illuminate
  - Staging – Reporting – Illuminate
- The data structures for the reporting database can be found in the Reporting Data Structures directory

Tableau Templates

We have provided sample Tableau templates. The SQL queries corresponding to the Tableau reports can be found in the git repository under the SQL Queries directory

Software Requirements

- Data Pipelines – Python version 3.6 or higher
- Tableau – version 2019.2 or higher