Data Warehouse Lab

Flynn Ramos

10/20/2019

Table of Contents

[Company Mission 3](#_Toc22489871)

[Perspective, Context, and Constraints 3](#_Toc22489872)

[Assumptions 3](#_Toc22489873)

[References 4](#_Toc22489874)

[Needs Assessment 4](#_Toc22489875)

[Automation 4](#_Toc22489876)

[Scheduler server 4](#_Toc22489877)

[Report Updating 4](#_Toc22489878)

[Extract, Transform, and Load 5](#_Toc22489879)

[Data acquisition 5](#_Toc22489880)

[Data Transformation and Cleansing 5](#_Toc22489881)

[Data Loading 5](#_Toc22489882)

[Warehouse 5](#_Toc22489883)

[Tables 5](#_Toc22489884)

[Storage buffer 6](#_Toc22489885)

[Database 6](#_Toc22489886)

[Analytics 6](#_Toc22489887)

[Data source 6](#_Toc22489888)

[Report Builder 6](#_Toc22489889)

[Dependencies 7](#_Toc22489890)

[OS 7](#_Toc22489891)

[Google Cloud 7](#_Toc22489892)

[Scripting 7](#_Toc22489893)

[User Groups 7](#_Toc22489894)

[Google Cloud 7](#_Toc22489895)

[Data Studio 8](#_Toc22489896)

[Security 8](#_Toc22489897)

[Authentication 8](#_Toc22489898)

[Task Scheduler 8](#_Toc22489899)

Company Mission

FoxTeam wishes to enter the Data industry but is having trouble with their data handling methods. They need a system that supports secure warehousing of Dungeons and Dragons related data. They expect to be able to easily access the database. Furthermore, they expect the system to build reports that are consistent and topical.

Perspective, Context, and Constraints

FoxTeam has received an opportunity to partner up with Wizards of the Coasts (WOTC) to develop a historical record of public characters built on their new web interface: DnDBeyond. That said they need to prove they are capable of such a task. As such WOTC hopes that FoxTeam can,

1. Prove that they’re able to handle a vast majority of user made data by loosing less than 5% of data upon loading into database.
2. Prove that they’re able to reduce maintenance costs by automating many of the systems functions.
3. Prove that they’re able to work in a data centric environment by providing clean and insightful reports

Assumptions

1. The server housing the scheduled tasks is assumed to already be in a secured environment
2. The data must be topical and remain within similar range

References

ERD – ERDPlusDiagram.PNG

Peer Review – CISY\_8503\_Peer\_Review\_Data\_warehouse

# **Needs Assessment**

# Automation

## Scheduler server

* 1. Uses Window’s Task Scheduler to run batch script (Script\_Scheduler.bat) to execute,
     1. Get\_DnD\_Raw\_From\_Repos.py
     2. Remove\_column.py
     3. CSV\_to\_JSON.py
     4. DeleteNullPlayers.ps1
     5. Clean\_and\_Make\_KVPairs.py

## Report Updating

* 1. Query connection between GCP BigQuery and GCP Data Studio

# Extract, Transform, and Load

## Data acquisition

* 1. source from GitHub (<https://github.com/oganm/dndstats/blob/master/docs/uniqueTable.tsv>)
  2. script Get\_DnD\_Raw\_From\_Repos.py

## Data Transformation and Cleansing

* 1. Script Remove\_column.py
  2. Script CSV\_to\_JSON.py
  3. Script DeleteNullPlayers.ps1
  4. Script Clean\_and\_Make\_KVPairs.py

## Data Loading

* 1. Manually transfer files to data bucket (data\_bucket\_e) on GCP: Storage
     1. Files to transfer are stored in DnD\_Data folder location

# Warehouse

## Tables

1. Fact table
   1. Players
2. Dimensions
   1. Boolean
      1. Classes-Bridge
      2. Feats-Bridge
      3. Spells-Bridge
      4. Weapons-Bridge
      5. Subclasses-Bridge
      6. Skills-Bridge
   2. Non-Boolean
      1. Alignments
      2. Races

## Storage buffer

* 1. Google Cloud Platform: Storage (data\_bucket\_e)

## Database

* 1. Star Schema (see ERD)
  2. Database Security (see User Groups below)
  3. Google Cloud Platform BigQuery

# Analytics

## Data source

* 1. Connection to BigQuery through Data Studio’s data sources

## Report Builder

* 1. Report access (see User Groups below)
  2. Google Cloud Platform: Data Studio

# Dependencies

## OS

* 1. Windows 10

## Google Cloud

* 1. Valid Gmail address with access to GCP solutions

## Scripting

* 1. Python (version 3.7)
  2. PowerShell (latest on windows 10)
  3. Batch (latest on windows 10)

# User Groups

### Google Cloud

* + 1. Granted through IAM Manager
       1. Validator Role
          1. BigQuery Data Viewer
          2. Storage Object Viewer
       2. Developer Role
          1. BigQuery Data Editor
          2. BigQuery Job User
          3. BigQuery User
       3. Loader Role
          1. Storage Object Admin
       4. Owner Role
          1. All permissions on project

### Data Studio

* + 1. Viewers given VIEW permissions
    2. Collaborators given VIEW and EDIT permissions

# Security

## Authentication

* 1. Database
     1. Valid Gmail added to IAM and given roles
  2. Reports
     1. Valid Gmail given view permissions manually through share function

## Task Scheduler

1. Data files deleted by batch script as last operation