LinkUp RAW 2.1 Improvements

January 2020

Document Overview

This document will outline the major differences and improvements to LinkUp's RAW product line in the 2.1 version. The intent is to provide high level information about changes and stand as a launching point for a more in depth conversation during migration from RAW 1.x to RAW 2.1. We fully expect more questions and to work through the migration with existing customers. Please also view the latest Raw 2.1 User Guide for a description of current file contents.

Job Records and Descriptions File Changes

Removal of Advertisement Jobs

We have removed all jobs where is_ad, linkup_hosted, linkup_hosted_historical are true from the job record files. We were always recommending clients to filter these jobs out before doing analysis, as these jobs are often fabricated and duplicated for SEO purposes. As such, we have removed them from the files, and subsequently removed those columns as they would now always be false.

Replacement of LinkUp Category with SOC/ONET

LinkUp's internal categorization system has been replaced by SOC/ONET codes in the job records files. Rather than a rudimentary set of string matching to determine the category of a job for LinkUp's website, we have replaced the category_name and category_id columns with an onet_occupation_code column which is determined by using an NLP service analyzing the full title and description of the job. We did this to better map to a standard categorization system and to improve the quality of our categorization of jobs.

Delete Date Column

We have added a delete date column to our job records to explicitly indicate when the job has been taken down by the employer. The date recorded is the first date we checked the employer website and did **not** find the url in question. If a job is considered to still be active this column will be blank. We now use this column in our analytics files to determine job activity.

Url Column

We have added the url column to the job records file. The url is the url of the page where the job title, description, and location were found. This column can be useful for investigating anomalies and analysis of a specific employers job posting habits.

Analytics File Changes

We have split what were known as the Analytics files into two distinct groups, Company Reference files and Core Analytics files. This was done to normalize the data for delivery, lowering the total file size needed for transfer. The files listed below are considered a replacement for Company Summary and Company Daily analytics files. Many other changes and improvements went along with this restructuring and are outlined below.

Company Reference Files

PIT Company Reference Files

This set of files is delivered daily as a differential and monthly as a full export, similar to the job records and descriptions files. It contains information about the company throughout time as LinkUp has tracked it. Things like company name, company url, and other employer specific data points are in this file. These have been removed from the Core Analytics files as these data points tend not to change as frequently as job counts.

PIT Company Scrape Log

This is a new addition to the product. It is provided as a daily log of yesterday’s activity and a monthly full log of all activity since the beginning of LinkUp’s data collection. This file allows a user to determine if a company was scraped on a particular day, and if any changes were made to that company’s scrape on a particular day. This can be useful in determining legitimacy of outliers.

FactSet Company Reference File

LinkUp has mapped our data to FactSet’s universe. This file is delivered daily as a full export of all companies. It contains that mapping as well as some financial identifier. LinkUp crawls FactSet’s entity hierarchy and flattens out ownership of private companies over time to provide a point in time look at a given LinkUp company’s publicly traded symbol, ISIN, CUSIP, and SEDOL.

Core Analytics Files

Core Company Analytics

This file contains daily company level counts of jobs on a daily basis. Each day the file contains the entire history of job counts dating back to 2007. The company id is provided in this file for easy joining to the above Company Reference files.

Core Ticker Analytics

This file contains daily stock ticker counts of jobs on a daily basis. Each day the file contains the entire history of job counts dating back to 2007. This uses the FactSet Company Reference data to determine the ticker over time. Only primary tickers are considered for this aggregation.

Data Handling And Operational Changes

24 Hour UTC Day Boundaries

RAW 2.1 daily job records and analytics files now respect UTC 24 hour boundaries. The previous version of RAW had some issues with the consistency of delivery and what was included in the daily job records differential, as it was operating off of our running scrape system data store. Now, we only included job data for scrapes that have fully completed within the UTC day when producing our daily content. This results in a clean cut off each day, rather than taking a measurement of an in flight system around 4 PM EST which leaves partially complete scrapes in the file.

Delivery Time of Daily Job Records and Descriptions

Due to the above change, we begin processing the daily files shortly after midnight UTC each day. Previously our daily files began processing at around 4 PM EST. This will change the time that files are available each day to be later, but include the full day worth of activity. All other files are delivered after the daily job records and descriptions pending processing time.

Delivery Platform

We have moved all of our new data files to a cloud based FTP service. Many of our clients had requirements of using FTP instead of AWS's S3, and our internal FTP server was hard to scale. As such we've moved all files to our new FTP located at files.linkup.com, a much more robust and reliable FTP solution. We also provide our own web based interface, Job Data Engine, for users looking to explore files with a browser and potentially expand their usage of LinkUp's data platform. Unfortunately S3 is not available for RAW 2.1 files, but we are looking to add more delivery options in the future.