Raw 1 to Raw 2 Improvements

Updated July 2020

Background

LinkUp first began offering the entirety of its database of job market data in mid-2014. The initial prototype was to simply export the raw data from the relevant databases into files suitable for compressing, transferring, and ingesting.

This arrangement was adequate for the first early adopters to research, evaluate, trial, and even use the data for production modeling and analysis. As the reach of this offering expanded, we adapted, evolved, expanded and improved the initial prototype to better suit the needs of an ever-broadening audience with an ever-increasing array of requirements and use cases.

As the dataset grew and as the number and variety of users grew, we had started to accumulate a list of items that could not be addressed without a major upgrade to the infrastructure, operations, formats, and delivery of the dataset. We designed and built the next generation of our LinkUp Raw product to address that set of issues. We now refer to this new version as Raw 2 and the original version as Raw 1.

Why Migrate

There are many benefits of migrating to our latest version of LinkUp Raw. Here are just a few of the benefits:

- Delivery is more timely and reliable
- The files have been broken up into smaller chunks, making the download times faster
- Extraneous data has been filtered out
- Reference data is much improved
- Scrape log that allows for correlating anomalies to interventions
- Future product improvements will be made to Raw 2 rather than Raw 1

More detail on the changes will be detailed below.

Job Records and Descriptions File Changes

Removal of Job Advertisements

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.

These extraneous “jobs” can be thought of more as entries that represent individual advertisements for jobs, but not the job postings themselves. The volume, timing, and content of these advertisements suffer from selection bias in that they are only LinkUp advertising clients (and only a subset, at that). They are also very idiosyncratic based on current conditions of the advertising environment as well as the current strategies of the particular recruitment advertising agencies administering the ad campaigns.

It has always been advised that users of the Raw 1 product filter out and ignore these jobs in their analyses. Omitting them from the set in the Raw 2 product saves everyone from unnecessary storage, transfer, compute, and complexity costs.

Here you can see a summary of the number of jobs in each set and a breakdown of what accounts for the differences (computed as of April 30th, 2020):

Total in 1:	269,086,580	
Total in 2:	144,223,827	(53.6% of v1)
In Both:	141,459,938	(98.0% of v2)
In 1 not 2:	127,626,641	(47.4% of v1)
Is_ad:	108,420,701	(40.2% of v1)
Linkup_hosted:	6,721,496	(2.5% of v1)
Linkup_hosted_historical:	11,026,750	(3.9% of v1)
Date Range:	2,813,446	(1.0% of v1)
is_ad has changed:	2,472,009	(1.5% of v1)
In 2 not 1 (is_ad has changed):	2,763,889	(1.9% of v2)

Replacement of Proprietary 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.

Addition of 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.

Addition of 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 employer's 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 identifiers. 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 ID'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 include 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 presented scaling challenges. As such, we've moved all files to our new FTP located at files.linkup.com, a 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.

Delivery of the Raw 2 files via S3 is in development and should be available by the end of the 2nd quarter of 2020.

Handling of Historical Changes

Periodically, we detect errors in jobs that have already been distributed. When the set of erroneous data points can be identified and isolated and are deemed to be material, we will correct the erroneous data points by updating or deleting job records in the historical set.

When this happens, we do not republish any of the impacted daily files, but subsequent full archives (published monthly) will reflect that change to the historical data. When this happens, we will publish a corrections notice on the Corrections page of the Resources section of the Job Data Engine application and send a notice via email to the account contact.

More Information

For more information on Raw 2, including more detailed product documentation, release notes, and tutorials, please see our GitHub page: https://github.com/LinkUp/Job-Market-Data.

Detailed File and Field Changes

Files from Raw 1 that are no longer included in Raw 2

Daily

Parent folder	File Path	File Name	Notes
aux	/daily/analytics/aux/YYYY/MM	LinkUpRaw_Aux_YYYY-MM-DD.tar.gz Categories_YYYY_MM_DD.csv.gz id_coverage_stats_YYYY-MM-DD.csv.gz	Mainly coverage documents, will not be included in raw 2
bucket	/daily/analytics/bucket/YYYY/MM	bucket_Daily_YYYY-MM-DD.csv.gz bucket_Daily_YYYY-MM-DD_d.csv.gz bucket_Daily_YYYY-MM-DD_d_1.csv.gz bucket_Daily_pit_YYYY-MM-DD.csv.gz bucket_Daily_pit_YYYY-MM-DD_d.csv.gz bucket_Monthly_YYYY-MM-DD.csv.gz bucket_Monthly_YYYY-MM-DD_d.csv.gz bucket_Monthly_YYYY-MM-DD_d_1.csv.gz bucket_Monthly_pit_YYYY-MM-DD.csv.gz bucket_Monthly_pit_YYYY-MM-DD_d.csv.gz	Bucket will not be offered in raw. Can aggregate by all features of the data set. Pre Aggregated files provide for company level and stock symbol level on a daily cadence. Replaced with ticker analytics, the clean version will be dropped. PIT will no longer be offered as an analytic file.
Macro	/daily/analytics/buckets/Macro/YYYY/MM	Macro_Daily_YYYY-MM-DD.csv.gz Macro_Daily_YYYY-MM-DD_d.csv.gz Macro_Monthly_YYYY-MM-DD.csv.gz Macro_Monthly_YYYY-MM-DD_d.csv.gz	Marco analytics not directly provided in raw 2 dataset. Can be created by aggregating data.
USMacro	/daily/analytics/buckets/USMacro/YYYY/MM	USMacro_Daily_YYYY-MM-DD.csv.gz USMacro_Daily_YYYY-MM-DD_d.csv.gz USMacro_Monthly_YYYY-MM-DD.csv.gz USMacro_Monthly_YYYY-MM-DD_d.csv.gz	Marco analytics not directly provided in raw 2 dataset. Can be created by aggregating data.

Macro_10000	/daily/analytics/buckets/USMacro_10000/YYYY/MM	USMacro_10000_Daily_YYYY-MM-DD.csv.gz USMacro_10000_Daily_YYYY-MM-DD_d.csv.gz USMacro_10000_Monthly_YYYY-MM-DD.csv.gz USMacro_10000_Monthly_YYYY-MM-DD_d.csv.gz	Marco analytics not directly provided in raw 2 dataset. Can be created by aggregating data.
USMacro_xbrks	/daily/analytics/buckets/USMacro_xbrks	USMacro_xbrks_Daily_YYYY-MM-DD.csv.gz USMacro_xbrks_Daily_YYYY-MM-DD_d.csv.gz	Marco analytics not directly provided in raw 2 dataset. Can be created by aggregating data.
state	/daily/analytics/buckets/state/YYYY/MM	State_Daily_YYYY-MM-DD.csv.gz State_Daily_YYYY-MM-DD_d.csv.gz State_Monthly_YYYY-MM-DD.csv.gz State_Monthly_YYYY-MM-DD_d.csv.gz	Not provided in raw 2. Can be created by aggregating raw records.
company	/daily/analytics/company/YYYY/MM	Company_Daily_YYYY-MM-DD.csv.gz Company_Daily_YYYY-MM-DD_d.csv.gz Company_Daily_YYYY-MM-DD_d_1.csv.gz Company_Summary_YYYY-MM-DD.csv.gz Company_Summary_YYYY-MM-DD_d.csv.gz company_daily_pit.csv.gz	Company analytics provided in raw 2. No PIT full files will be provided. There will just be one company file with the following analytics: active, created, deleted, duration.This file will correct for drift. File Path - /JDE/Standard/MarketReports/Reports/Core_Ticker_Analytics
company_daily	/daily/analytics/company_daily/YYYY/MM	Company_Daily_YYYY-MM-DD.csv.gz Company_Daily_YYYY-MM-DD_d.csv.gz Company_Daily_YYYY-MM-DD_d_1.csv.gz	Company analytics provided in raw 2. No PIT full files will be provided. There will just be one company file with the following analytics: active, created, deleted, duration.This file will correct for drift. File Path -

			/JDE/Standard/MarketReports/Reports/Core_Ticker_Analytics
company_daily_b	/daily/analytics/company_daily_b/YYYY/MM	empty	N/A
company_summary	/daily/analytics/company_summary/YYYY/MM	Company_Summary_YYYY-MM-DD.csv.gz Company_Summary_YYYY-MM-DD_d.csv.gz	No similar file will be provided in raw 2
hash	/daily/analytics/hash/YYYY/MM	Hash_YY-MM-DD Hash_YY-MM-DD	onet code on job records Inferred delete date no longer in this file
raw	/daily/raw/2020/06	linkup_raw_daily_YYYY-MM-DD	File is now /daily/raw/YYYY/MM

Archive

File	File Path	File Name	Notes
analytics	/archive/analytics/YYYY/MM	Daily_LinkUpRaw_Analytics_YYYY-MM-DD.tar.gz	Archived package of all analytics will not be available in raw 2. The 2 analytics files will have complete history.

Files that have changed from Raw 1 to Raw 2

Parent folder	File Path	File Name	changes
jobs	/archive/jobs/YYYY/MM	linkup_raw_YYYY-MM-DD.tar.gz	Following columns will not be included in records file - category_id - Category_name - Is_ad - linkup_hosted - Linkup_hosted_historical New Fields added to record file - Onet_occupation_code - url
raw	/daily/raw/YYYY/MM	linkup_raw_daily_YYYY-MM-DD	Following columns will not be included in records file - category_id - Category_name - Is_ad - linkup_hosted - Linkup_hosted_historical New Fields add to record file - Onet_occupation_code - url
descriptions	/archive/descriptions/YYY/MM	linkup_raw_YYYY-MM-DD.tar.gz	

New Files in Raw 2 that were not in Raw 1

RAW2 File Name	Notes	File Path
FS Daily Reference	This file is used to roll company_id up to tradable stock symbols	/JDE/Standard/Feeds/FS Daily Company Reference/YYYY/MM
Raw Daily Company Scrape Log	This file is used to flag if a scrape has changed. If a scrape has changed that likely means a scrape break took place. This is the daily scrape file, can be inserted into the full	/JDE/Standard/Feeds/Raw Daily Company Scrape Log/YYYY/MM
Raw Full Company Scrape Log	This file is used to flag if a scrape has changed. If a scrape has changed that likely means a scrape break took place.	/JDE/Standard/Feeds/Raw Full Company Scrape Log/YYYY/MM
Raw Full Company PIT reference	This file is used to record data attributes of the companies we scrape. This file includes things like company name and website.	/JDE/Standard/Feeds/Raw Full PIT Company Reference/YYYY/MM
Raw Daily Company PIT reference	This file is used to record data attributes of the companies we scrape. This file includes things like company name and website. This is the Daily file.	/JDE/Standard/Feeds/Raw Daily PIT Company Reference/YYYY/MM
Raw core company analytics	Daily company analytics, each file is the full analytics for all companies	/JDE/Standard/MarketReports/Reports/Core_Company_Analytics
Raw core ticker analytics	Daily stock ticker analytics, each	/JDE/Standard/MarketReports/Reports/Core_Ticker_Analytics

	file is the full analytics for a ticker	