# RELEASE UPDATES and Changes

|  |  |
| --- | --- |
| **1/3/2022** | **We have officially released the “active\_log”. active\_log is a point in time log of each jobid that is active on any given "thrudate", which in this case is the equivalent of the "jobsdate" in the analyticsCore files analytics. This file basically fills all the gaps in the jobs\_log (filling in between days a job might not have been scraped), as well as forward filling for currently active jobs to the marginal "current" date. This can be used to create your own analytics quite easily.**  **IMPACT**  **You will see a new “folder” called “active\_log” (you may have already seen). And you will have full access to it starting tomorrow (1/4/2022). This file just contains just jobid and a date.**  **What will this do for me?**  **This file is a parquet that is partitioned by “thrudate”. Basically, if you select a specific “thrudate” partition, you will see all the “jobid”s that are active on that date – thus the number of active jobs on that date. With this, you can join the active\_log to any number of reference data items (like we do for ticker, city, state, zip, you name it) to GROUPBY and get job counts by thrudate for the reference data item. It saves you the step of modifying the jobs\_log when you want to create your own analytics should you choose not to use ours. Also, by looking at the changes day to day, you can create the jobscreated, jobsremoved, and calculate the duration\_active and duration\_closed.**  **Please see the** [**Glossary**](https://github.com/SmartMarketData/LinkUp_Enhanced_Documentation/blob/master/LU_Enhanced_Glossary.xlsx) **for Schema details, and expect to have full access on 1/4/2022.** |
| **10/7/2021** | **We had a production issue last night which incorrectly “missed” the job posts that were deleted on 2021-10-06.**  **IMPACT**  This error made it so that all companies showed zero “jobsremoved” (deleted job posts). This impacted several files which we recommend you REPLACE if you have already ingested the Gold Copy files for yesterday. They are:   1. Download the previously missing ***jobs\_log/scrape\_date=2021-10-06/addremoveflag=-1*** partition 2. Replace all the ***analyticsCore/…/jobsdate=2021-10-06*** partitions for the various aggregates 3. Replace the ***analyticsEnhancedJob/created\_pit=2021-10-06/analyticType=reposts\_seo*** partition 4. Replace the ***analyticsEnhancedJob/created\_pit=2021-10-06/analyticType=url*** partition 5. Replace the ***analyticsEnhancedJob/created\_pit=2021-10-06/analyticType=socCode2010*** partition 6. Replace the ***analyticsEnhancedJob/created\_pit=2021-10-06/analyticType=reposts\_seo*** partition 7. Replace the ***active\_log/thrudate=2021-10-06*** partition   There are no other changes, no history was changed. No historical files were changed. You do NOT need to redownload everything, just replace those file partitions above with the new “fixed” ones.  We are sorry for the issue, and do not expect any future errors.  Please let us know if you have any questions. Thanks you. SMD Operations… |
| **7/13/2021** | **RELEASE of Work From Home Job Post Flag**  We will be releasing our “Work From Home” job post flag today.  This flag uses our proprietary NLP process to evaluate the job post “title” and “description” to determine whether a job post is a “remote” or “work from home” job.  We specifically look for work from home, WFH , remote work, online work, remote eligible, home based, telecommuting and variants.  This new flag is available in the ***analyticsEnhancedJob/created\_pit=\*/analyticType=word\_workfromhome*** parquet partition.  You will find that this parquet has a field “value”.  For this analytic, the “value” is the score we assign to the flag.  This value is based on a bunch of thing, primarily matches.  So a 12 would be high and looks like 29 currently the highest.  You can use this “value” to fine tune your internal models/analytics based on your tolerance for accuracy.  1 would be least confident, 29 highest in this case.  There is theoretically no upper limit for the score, although size of descriptions will put a top.   Here is the distribution of scores:    Here is a quick chart of the work from home job posts that were created each day.  Coming soon we will be creating “**analytics\_advanced**” in the **analyticsCore** parquet, which will incorporate aggregate level analytics for our part time jobs, as well as work from home jobs, among other things.  So stay tuned for that announcement this summer.  If you have other broader level flags you would like us to process, let us know!    **IMPACT**  All you will see is a “new” partition in the **analyticsEnhancedJob** parquet called “***analytcType=word\_workfromhome***”.  No existing files or schemas have changed.  You can consume this new flag, or ignore it based on your needs. |
| **7/5/2021** | Corrected periodic duplicate records in the “**reference**” parquet, including ***reftype=ticker\_ric*** and ***reftype=MatchScoreFlag*** |
| **7/5/2021** | Removed “nan” jobid records from **the jobs\_log/scrape\_date=2015-03-14/addremoveflag=0** parquet partition. This replaced parquet file ***e00a6871b6ef4128a91cc3d47e34cfa4.parquet*** with **5ce688bd3ae44477a8cc0f6e63746321.parquet**. No change in valid records, no impact on analytics. |
| **6/29/2021** | Hi all, per our previous communication, we have completed the update of our “Part Time” job flag analytics.  **IMPACT:**  You will see an overall increase of ~871K new flagged jobs. We originally flagged ~16.8M job posts as part-time jobs, we now flag 17.6M. These added part-time flags are the result of us adding ~5.25M additional job descriptions, per the previous communication. You can see the impact is primarily felt starting 2020-08-01 in the below chart. The orange line is the new flag counts, the blue line the old flag counts.  You will want to RELOAD the entire ***/analyticsEnhancedJob/created\_pit=\*/analyticType=word\_parttime*** parquet partition. Then you can resume just loading the incremental point in time created\_pit that is available daily.    This update does not restate any of the historical aggregate analytics (analyticsCore) or raw data (jobs\_base and jobs\_log).  We will be releasing our “Work From Home” flags over the weekend as well!  In addition, we expect to release our first set of aggregate advanced analytics during July, the first of which will be the “part-time” and “work from home” job post counts at the aggregate level (company\_id, ticker, sedol, country etc…).  If you have any questions, please reach out. Thank You. operations |
| **6/23/2021** | The Descriptions Update has been completed.  Attached is a count of the new added Descriptions, by company\_id.  We recommend that if you use this file, and are not directly syncing, that you delete your existing version, and reload the new updated version.  The parquet file is ***smd-lu/descriptions.***  If you are creating your own analytics based on the Descriptions data, we would recommend that you re-run them to include the missing descriptions.  We will now be processing the additional jobs for our “part-time” job indicator.  Once complete, we will replace the existing “analyticType=word\_parttime” partition under the “analyticsEnhancedJob” parquet.  We will notify you when this is complete so that you can reload this specific flag.  We are targeting 2021-07-01 for these additional analytics.  Thanks all.  Operations… |
| **6/22/2021** | **Notice of addition of Job Descriptions that were previously missed**  Tonight we will be adding approximately 5.25M additional historical job descriptions to the “descriptions” parquet file.  The LinkUp team discovered a bug which prevented a portion of the Descriptions to be missed starting on 2020-08-01.  The team fixed the bug and have released the additional descriptions.  Impact:  We will be republishing the ***descriptions*** file this evening.  We recommend that if you use this file, and are not directly syncing, that you delete your existing version, and reload the new updated version.  Subsequent to this, we will be processing the additional jobs for our “part-time” job indicator.  Once complete, we will replace the existing “analyticType=word\_parttime” partition under the “analyticsEnhancedJob” parquet file.  We will notify you when this is complete so that you can reload this specific flag.  We are targeting 2021-07-01 for these additional analytics.  If you are creating your own analytics based on the Descriptions data, we would recommend that you re-run them to include the missing descriptions.  There are no changes to any of the historical analytics, raw jobs data (job\_log, job\_base), or the core aggregate analytics (analyticsCore).  If you have any questions, please reach out.  Thank you.  lg |
| **4/26/2021** | **ENHANCEMENT – reposts\_seo**  We have added a NEW analyticsEnhancedJob hash level analytic – ***analyticType=reposts\_seo***.  “**reposts\_seo**” are hash that we flag as "reposts" that we believe may be the result of 1) companies removing jobs, and immediately replacing them so they look "new", and therefore appear at the top of search engine requests, or occassionally 2) a potential scraper break. **Reposts\_seo** are a subset of ALL reposts, but represent around 50% of all reposts (on 4/26/2021 there were ~23M repost\_seo in the data set, versus ~43M total reposts). Using these flagged repost hash to adjust the point in time data will reduce some of the volatility in the point in time data, but is NOT what you would have seen on a given day. Using these to adjust/smooth the back data, creates a "lag" at the company\_id level to the "last" scrape\_date. If a company is scraped daily, then the lag is 1 day. If the last scrape\_date was 2 days ago, the lag is 2 days for that company. IF the company last scraped 1 month ago, then a 1 month lag etc...This lag occurs because we don't know if a job is a "1 cycle" repost until we see the latest/marginal scrape. The "value" field is filled with the "jobid" of the reposted job. You can use these flagged jobs to make your own adjustments, or to verify the adjusted analytics we will be providing the summer of 2021. Remember also that you can see all the reposts in the “jobs\_base” where repostflag==1. |
| **3/27/2021** | **FIX**  We have removed duplicated ‘hash’ records in the analyticsEnhancedJob ‘socCode2010’,’word\_parttime’,’url’,’and ‘badhash’ which were erronesously included during some multiprocessing. These were just duplicate records, and no records are changed. We have put in place process to check for these at production.  You will see the deletion of some parquet files, and replacement with new parquet files for the affected created\_date partitions. |
| **3/23/2021** | **ENHANCEMENT**  We have enhanced the job ‘descriptions’ raw data.  First, we have identified and removed all of the email addresses in the Description text as part of our effort to clean up the primary source of Personal Identifiable Information in the LinkUp Job Post “descriptions” dataset. We have replaced all email addresses with [cleaned@emailaddress.com](mailto:cleaned@emailaddress.com). As of 3/22/2021 there are roughly 6,310,775 email addresses that are now cleansed in the data set.  We have also further cleansed the RAW text, removing white space, as well as various other “noise” from the scrape process – primarily rogue html.  Finally, we had some duplicated ‘hash’ in the panel that we removed.  None of these changes affect any analytics. And no descriptions were removed or added.  The enhanced descriptions should be up by 0:00 UTC 3/25/2021. If you consume the ‘descriptions’ parquet files, we recommend that you re-download the entire parquet. |
| **1/5/2021** | **CHANGE**  We discovered an anomaly in the analytics which was created when a company removed ALL of their jobs (so we scraped and we found no jobs posted). The “jobsremoved” was not getting captured correctly. We have fixed this issue, and rerun the analytics. The primary “jobsactive”, as well as the other “core” analytics remained accurate, the only impact was on the “jobsremoved”.  In addition we identified an anomaly in our processing of “reposted” jobs, which created some “duplicates”. We identified this anomaly and corrected it as well. This had a very small impact on the “raw” jobs data files including the jobs\_base and jobs\_log.  The impact is insignificant at the Macro level, and only impacts a small number of jobs and company\_ids, many of which are for private companies. Our pre-fix analysis shows that the “bug” impacted 535 total company\_ids, 148 created dates, and 6,244 jobs.  The first date affected was 2016-06-08 with 13 jobs.  The last date was 2020-04-30 with 4,062 jobs by 469 company\_ids.  We process this raw data in multiple subprocesses.  It appears that the likely culprit was a periodic “overlap” of multiple processes regarding the impacted dates.  So it looks like the issue is fairly insignificant affecting only 0.0000334 of the jobs.  Due to these fixes, we will be swapping out the old jobs\_log, jobs\_base, and the “analyticsCore” files.  **ENHANCEMENT**  Finally, we are happy to also now provide some additional aggregate analytics in the “analyticsCore”. We are adding the SOC code aggregates, broken down as follows from the [Aggregate Analytics Codes](LinkUp_Enhanced_Documentation/LU_Enhanced_Aggregate_Analytics_Codes.xlsx%20at%20master%20·%20SmartMarketData/LinkUp_Enhanced_Documentation%20·%20GitHub):   |  |  | | --- | --- | | SOC Major | 00011700000001 | | SOC Minor | 00011710000001 | | SOC Broad | 00011720000001 | | SOC Detailed | 00011730000001 | |
| **12/16/2020** | Added actionType ‘scrape\_changed’ to the scrapelog. |
| **11/17/2020** | Added reftype(s) ‘employee\_full\_time’ and ‘employee\_part\_time’ to the reference file. These are gathered from company SEC filings and are located in partition refsource=’SMD’. |
| **11/15/2020** | Descriptions – we have reposted a new “descriptions” parquet where we have adjusted the “created\_pit” partition to the first day the “hash” job description was posted. In the previous version the created\_pit was synced with the jobs\_base created\_pit, which had the effect of “back\_dating” the actual pit time you actually would have seen the description for the first time . In other words, sometimes jobs are posted, and the first time the description for the job is posted is some day after. This change will now accurately reflect the point in time you would have first seen the description. Note that the actual descriptions are the same, this just adjusted a small % of the created\_pit’s used for partitioning. The net effect should be now just 1 file in each created\_pit partition… |
| **9/11/2020** | Macro Job level "part-time" jobs flagged and scored. See //analyticsEnhancedJob --> analyticType=word\_parttime |
| **9/11/2020** | Macro Job level "bad hash" flagged. These are jobs that LinkUp determined were invalid, and retroactively removed from the full data set. See See //analyticsEnhancedJob --> analyticType=badhash |
| **9/9/2020** | Macro Zip Code level aggregate analytics released. See //analyticsCore// --> aggregate code=00011130000001 |
| **9/3/2020** | Macro State level aggregate analytics released. See //analyticsCore// --> aggregate code=00011110000001 |
| **8/31/2020** | Macro Country level aggregate analytics released. See //analyticsCore// --> aggregate code=00011100000001 |
| **8/10/2020** | LinkUp Raw2 Enhanced - Gold Copy released |
|  |  |
|  |  |