**LinkUp RAW 2 Enhanced – Gold Copy**

**Getting Started**

Table of Contents

[LinkUp RAW Jobs Data 2](#_Toc80256724)

[Summary 2](#_Toc80256725)

[Files Needed 4](#_Toc80256726)

[The Files 5](#_Toc80256727)

[jobs\_base 5](#_Toc80256728)

[jobs\_log 6](#_Toc80256729)

[scrapelog 7](#_Toc80256730)

[reference 8](#_Toc80256731)

[descriptions 10](#_Toc80256732)

[auxiliary 10](#_Toc80256733)

[analytics\_core 11](#_Toc80256734)

[The “Smart Code” 11](#_Toc80256735)

[Python Examples 14](#_Toc80256736)

[analyticsEnhancedJob 17](#_Toc80256737)

[File Glossary & Delivery Timing 22](#_Toc80256738)

[File Sizes 24](#_Toc80256739)

[AWS S3 access 24](#_Toc80256740)

[Details for implementing the “role” model 24](#_Toc80256741)

[AWS IAM Policy to Grant Access to LinkUp Data Files in S3 24](#_Toc80256742)

[Attach the AWS IAM Policy to Grant Access to LinkUp Data Files in S3 to an IAM user 25](#_Toc80256743)

[Attach the AWS IAM Policy to Grant Access to LinkUp Data Files in S3 to an IAM group 25](#_Toc80256744)

[Attach the AWS IAM Policy to Grant Access to LinkUp Data Files in S3 to an IAM role 26](#_Toc80256745)

[FAQs 26](#_Toc80256746)

[8) PII – Personally Identifiable Information 29](#_Toc80256747)

[9) What is the “main company\_id” for a ticker? 29](#_Toc80256748)

# LinkUp RAW Jobs Data

The LinkUp Raw Jobs data is a database of company level job posts, gathered directly from company career portals like “ibm.com/careers”. We do not capture jobs from “aggregators” like Indeed. This makes our raw data clean, with no duplicates and no stale data – among other things.

The data consists of raw data files, as well as analytics. The RAW data consists of job level meta data, and descriptions.

The engine behind LinkUp’s Job Search platform is driven by a very sophisticated and complex technology infrastructure that systematically “spiders” corporate web sites to gather information about open job postings.

While scraping an individual website once is not a technological feat today, doing it daily for greater than 55,000 companies is. LinkUp has spent years developing technology to deliver the most reliable, accurate and timely jobs data in the industry.

The jobs data history extends back to 2007-08-03, and covers some 55,000+ companies. The data is captured, analytics processed, and delivered daily. We started capturing the descriptions data in 2014.

# Summary

The ***LinkUp RAW Enhanced Gold Copy (LUGC)*** is a job level, point in time version of the LinkUp RAW Jobs (LU) data set. The primary advantages of LUGC are:

1. **Greatly simplified ETL**
2. **Jobs, not company\_ids**. The “job” becomes the atomic level of the raw data and analytics in LUGC, versus the “company\_id’ in LU. This allows for extremely detailed, flexible, and accurate analytics.
3. **Pure point in time (PIT).** With LUGC, we have created a PURE point in time representation of the RAW data, ***as you would have seen it on each day***, back to the inception of the data on 2007-08-03. With LU, you only have RAW data starting in the middle of 2018, and analytics starting in November of 2019. Meaning you are seeing the data as you would have seen it in 2018/2019 respectively, and you would have to build your own “near” point in time representations on a day to day basis. LU is NOT point in time.
4. **Normalized and cleansed.** LUGC corrects:
   1. **inaccurate “deleted” dates in LU**, affecting some 20+million hash. These inaccurate dates created look ahead bias as well as analytics “mismatch” where the logical “truth” that active jobs “today” should = active jobs “yesterday” + created jobs “today” – deleted jobs “today” was broken
   2. **removal of NULL “hash”**
   3. **correction of duplicate “hash” that have multiple company\_ids**
   4. **point in time removal of “bad” hash on the day LU defined them as “bad”**. Not waiting for the monthly ”archive” fix
   5. **point in time removal of jobs when a “company\_id scraper” is decommissioned**
   6. **correction/normalization of country and state values**
   7. **further cleansing of the “title” field**
   8. **normalized timestamps**
   9. **eliminate files with different encoding** (LU daily files shift from ‘utf8’ to ISO 8859-1 for 2 months in 2018)

***and LUGC “normalizes”:***

* **“created” dates.** LUGC has a “created\_pit” (this is the actual point in time production “created” date. LU assigns the created at the actual time the job was scraped, all good. BUT, they then assign the “last\_checked” date when the company\_id level scrape ended – which determines the date the information is published to users. This creates situations where in the daily scrape cycle, the job was created on 1 day, but “checked” on another. This creates many problems PIT, and creates a scenario where historical analytics are “changed” from day to day. With LUGC, jobs and analytics are perfectly synced with the production date – the day a client would have actually seen the information.

1. **Fixed and Enhanced the scrapelog**. The LU scrapelog was incomplete. You had many company\_ids that you know were scraped on certain days, but no record of the scrape. In addition, the LU scrapelog did not capture when company\_id “scrapers” were decommissioned. The new LUGC scrapelog represents all possible known scrapes for ALL company\_id’s throughout the entire history back to 2007-08-03.
2. **Accurately captures job “reposts”.** When a company would “remove” a job, then sometime later “repost” it, LU would just change the “deleted” date from NULL to the date of the initial “removal”, then back to NULL when the company “reposted”. This is problematic in several ways. First, you lose any daily point in time information about the actual job posting. From your perspective today, the job was ALWAYS posted, when in fact we know it was removed for a period. In addition, the “deleted” date, like the “created” date, should be immutable. It should not change from NULL, to a date, to NULL again. Finally, the analytics created from this are continually restated upwards, overstating historical data, resulting in an “upward drift” of active job counts.
3. **Cutting edge usage of .parquet file format**. Smaller footprint, flexible and fast consumption, verified schemas, and “Smart” partitions.
4. **Greatly simplified file naming topology and organization**. No need to worry about daily or archive csv’s or xml’s., or multiple locations. LUGC has simple straightforward naming conventions.
5. **Greatly reduced file storage footprint**.
6. **Elegantly designed files allow for scaling in new analytics without changing schemas**
7. **Reference data consolidated into 1 file.**
8. **Greatly enhanced number of Aggregate Analytics (like core analytics by country, city, state, zip, all levels of SOC code, all levels of NAICS, and many more)**
9. **Job level enhanced analytics like “part time” and “full time” among many others**
10. **Addition of EnhancedAnalytics such as “Closed Duration” and “Normalized Active Jobs” among many others to be added**
11. **Job “Descriptions” cleansed, including email address PII removed**

# Files Needed

Before we continue, you can find the LU\_Enhnced\_Glossary.xlsx and the LU\_Enhanced\_Aggregate\_Analytics\_Codes.xlsx, as well as other documentation and information here:

<https://github.com/SmartMarketData/LinkUp_Enhanced_Documentation>

The files are designed to be simple, and the topology has been simplified greatly. Basically, when you look at the data repository, this is what you see:



So ***what files do you need?*** Depends.

**If you want to build your own analytics, you will want:**

1. jobs\_base
2. jobs\_log
3. reference
4. scrapelog (not needed but can be used to verify data and look for scrape changes that may be responsible for quirky analytics)

**If you want to process the descriptions on your own with NLP, you will want:**

* jobs\_base
* descriptions
* jobs\_log (needed to further create analytics)
* reference (needed to create certain types of aggregates like “ticker” level)

*\*Note that SmartMarketData offers bespoke NLP work/analysis/analytics for any data set…*

**If you want just our analytics, you only need:**

* analyticsCore

**If you want to slice certain jobs out of the total set of jobs, you will need:**

* analyticsEnhancedJob

# The Files

jobs\_base **– ~9GB, adding <~10MB per day.**

This is an enhanced “jobs” table, which is just the immutable data for a given hash, with some important modifications to the LU standard data. This file is structured with the following schema:

* **jobid** – the jobid is a “smart” identifier. The first 8 digits are the created date, point in time, meaning the date that the client would have seen that the job was created the actual production date (see below “created\_pit”). The next 5 digits are the company\_id. The final 5 digits is the specific number of the job that was created that day for that company\_id. For example 202006090000100001 would be the jobid for Target for the first job that was created on 2020-06-09. While “hash” is random, this identifier has information and therefore value, flexibility and simplicity.
* **repostflag** – 0 or 1 based on if a repost or not
* **company\_id**
* **title**
* **city**
* **state**
* **zip**
* **country**
* **created** – we still give the original LU created as it is the “actual” scrape time vs. the “production” scrape time
* **parenthash** (this is the parent hash for the jobid. With this we can deal perfectly with all reposts).
* **created\_pit** (this is the actual PIT production created date, found also in the jobid, and fixes the “created” and the production mismatch issue.)



This and all the other files can be consumed directly from either your locally downloaded version, or from our S3 bucket, or your sync’ed AWS bucket.

To consume this file from a local source into a python dataframe for example, all you need to do is:



Where filedir is the local directory and parquet directory, for instance “c:\\lugc\\jobs\_base”

This will give you the entire jobs\_base file. No need to get monthly archives etc. And if you want to just grab a specific day, like the most current day, you can do this:



And, of course, you can grab multiple days, date ranges etc.

Ideally, you would have your “version” of LUGC synced with our AWS bucket. If you want to consume the data directly though (not download it locally) from either our bucket or yours, you can do this:



Where you would add your AWS credentials to the ‘key’ and ‘secret’ args, and the bucket…

The jobs\_base file becomes the “base” to the whole data set. It is point in time, and efficiently partitioned by the jobid’s “created\_pit”. You can consume it all, or just the dates you want, and never need to duplicate the storage or fields in your data repository. There is no monthly archive. No need for “deltas”. No need for separate point in time files as they are by nature PIT. This file becomes the base “jobid” reference table, replacing the LU “jobs” file.

jobs\_log **– 17GB, adding <~10MB per day.**

With the corrected/enhanced company level scrapelog (see below), and the jobs\_base as a foundation, we have created a true hash level point in time log for the entire history of the data. You can now look at 2007-08-03, and see the “daily” view on that date. The result though is a perfect PIT dataset, at the atomic “hash” level. This dataset is approaching 5 BILLION records! BUT, because of how we have constructed it, it requires only ~14GB of storage space, and is flexibly and quickly consumable vs. LinkUp’s current .csv and .xml methods. No archives, dailies, pit, or deltas needed. Pull the whole file up to date, or just select the scrape\_date (s) you want. Here are the fields:

* **jobid** – see above in jobs\_base
* **scrape\_date** – the PIT scrape date for the job
* **addremoveflag** - The addremoveflag is SMART!!! It can be either 1,0,-1, where 1 means the jobid was created, 0 means it was “last\_checked”, -1 means it was deleted.

This file is partitioned by “scrape\_date”, so you can pull all the jobs that were scraped or deleted on a specific date. It is also partitioned by the “addremoveflag”. So you can pull all “created” jobs where addremoveflag=1, or all “checked” jobs for addremoveflag=0, or all “deleted” jobs for addremoveflag=-1.

Like jobs\_base above, the parquet files can be consumed directly and efficiently.



scrapelog **– 220MB, adding <~50KB per day**

The scrapelog completes and improves the current LU Raw\_company\_scrape\_log\_daily\_yyyy-mm-dd.csv.gz.

We have created an enhanced “scrapelog”, which eliminates ALL of the gaps in the previous log and provides a complete historical base of company level scrapes. The scrapelog adjusts the deleted dates to point in time, so they reflect reality as you would have known it on any given day. We have also incorporated the dates that company level scrapes were decommissioned.

The scrapelog is quite efficient with fields:

* **company\_id**
* **actionDate**
* **actionType**. (which now can be “scrape”, “inferred scrape”, “scrapeadded”, “scrapedeleted”, “scrapechanged”)



reference **– ~13MB**

This is the consolidated reference data file. It contains LinkUp reference data, like the “company\_name” associated with the “company\_id”. It also contains the market data reference items like tickers, sedols, etc., as well as broader identifiers like LEI. The file is also parquet, and has the following schema:

* **company\_id**
* **refsource** – the source of the data
* **refsource\_id** – the primary identifier used by the refsource
* **reftype** – the type of reference data – i.e. URL, or company\_name, sedol, or ticker etc.
* **start\_date** – the start date of the mapping
* **end\_date** – the end date of the mapping
* **value** – the value (i.e. ‘AAPL’)
* **thrudate** – the date that the data was added to the file

This parquet is partitioned by ***“refsource”*** (i.e. Refinitiv/Factset/LinkUp/SmartMarketData etc.) as well as ***“reftype”*** (i.e. ticker, cusip, sedol, LEI, PermID etc.) for a fast and easy way to grab all the data at once, or just what you want (like the Factset stock\_ticker for instance).

The currently available refsource/reftype’s are:

* **Factset**
  + stock\_ticker\_primary
  + stock\_ticker
  + sedol\_primary (requires authorization)
  + sedol (requires authorization)
  + isin\_primary (requires authorization)
  + isin (requires authorization)
  + factset\_entity\_id (requires authorization)
  + cusip\_primary (requires authorization)
  + cusip (requires authorization)
* **LinkUp**
  + open\_perm\_id
  + naics\_code
  + lei
  + company\_url
  + company\_name
* **SMD**
  + URL
  + TRBC\_IG (requires authorization)
  + TRBC\_I (requires authorization)
  + TRBC\_ES (requires authorization)
  + TRBC\_BS (requires authorization)
  + ticker\_ric (requires authorization)
  + OrgID
  + OpenPermID
  + NAICS1
  + MatchScoreFlag
  + Match Score
  + ManualCheckFlag
  + LEI
  + CIK



descriptions **– 102GB, growing ~80MB per day**

These are the job descriptions from each job post. They are unstructured text. LU started to capture these in 2014, although jobs that were present before 2014 and active when we started capturing the descriptions will show up as well. You will see descriptions all the way back to 2007 for evergreen type jobs. In general, starting in 2014, most jobs have an associated description.

The descriptions parquet schema is:

* **hash** – joins to the parenthash in the jobs\_base
* **description** – the unstructured text
* **hash1** – this is so you can iterate through this large data set easily in the file/partition. Simplifies processing
* **created\_pit** – this is so you can pull specific days or just add a new day. Simplifies processing

Again, you can grab the entire set of descriptions, or you can grab by the first character of the hash (0-9,a-f), or by the “created\_pit” that the ‘hash’ was originally created.



auxiliary – **6MB – grows slightly periodically**

This directory contains auxiliary tables, typically from external sources as “helper” tables. For instance, the ‘/smd-lu/auxiliary/soccode/soc\_2010’ directory parquet contains all of the 2010 soc code descriptions. The available files are:

* /smd-lu/auxiliary/soccode/soc\_2010 – the soc code description, partitioned by “code”. Pull the whole file, or a specific code(s).
* /smd-lu/auxiliary/soccode/soc\_2018 – the soc code description, partitioned by “code”. Pull the whole file, or a specific code(s). (2018 SOC codes are coming in the future)
* /smd-lu/auxiliary/naics – naics descriptions coming soon
* /smd-lu/auxiliary/trbc – trbc descriptions coming soon

The schema for the soccode tables is:

* title
* definition
* code

analytics\_core – **each Aggregate can be from 20-1GB depending on the aggregate**

The analytics\_core is our “core” analytics, created for any “aggregate”. An aggregate is a subset of jobs. For instance “Macro” is all jobs. “USMacro” is just jobs with country=’USA’. An infinite number of aggregates can exist.

### The “Smart Code”

The “aggregates” are organized by a “smart code”. You simply look at the [table description](https://github.com/lgreen4/LinkUp_Enhanced_Documentation/blob/master/LU_Enhanced_Aggregate_Analytics_Codes.xlsx), determine the aggregate you want and it’s “smart code”, and extract that “smart code” to get all the available aggregate data (download locally, sync to your AWS S3 bucket, or extract directly from our source AWS S3 bucket at /smd-lu/analyticsCore/{smart code}/analytics\_core . The code is a 14 digit code such as:

00011010000001

The first four digits are the “main aggregate id”, such as:

* Macro=0001
* USMacro10000=0002
* Word\_Parttime=1001
* Many more to come!

The next digit is the “level”. This can be:

* Point in time = 1,
* Non-Point in time = 2
* Point in time – cleansed = 3
* Non-Point in time cleansed = 4
* Others possible

The next 2 digits are the “bucket” code. This is what the “blevel” field is in the aggregate panel. This can be:

* macro=00
* company\_id=01
* ticker=02
* ticker\_fs=03
* sedol\_fs=04
* country=10
* state=11
* city=12
* zip=13
* naics\_s=30
* naics\_ss=31
* naics\_ig=32
* naics\_i=33
* naics\_si=34
* trbc\_es=35
* trbc\_bs=36
* trbc\_ig=37
* trbc\_i=38
* soc\_major=70
* soc\_minor=71
* soc\_broad=72
* soc\_detailed=73

The next 2 digits are the “sub-bucket” code. For instance if you want to see all soc\_major(s) by company\_id, you would have 7001. These are be the same codes as the “bucket” code above.

The next digit is the “top” code. This is where we can assign the number of “top” we want in the aggregate. For instance, the “top 10,000”, or “top 50”. These codes are:

* 1=1
* 2=10
* 3=50
* 4=100
* 5=500
* 6=1000
* 7=2000
* 8=3000
* 9=5000
* 0=10000

The next 3 digits are the “client code”, and are assigned for specific clients who want to keep proprietary analytics. The default is ‘000’ which means the aggregate is not client specific.

The final digit is the “period” code. These can be:

* 1=daily
* 2=weekly
* 3=monthly
* 4=quarterly
* 5=semi-annually
* 6=annually

Some common aggregates:

* Macro
* Country
* State
* Zip
* Company\_id
* Ticker
* Ticker\_fs
* Sedol\_fs
* XYZ Hedgefund’s universe
* Detailed SOC Code
* GICS Sector
* TRBC Industry
* NAICS Sub Industry

You name it, we can create it. Just drop a note to [operations@smartmarketdata.com](mailto:operations@smartmarketdata.com) if you have an idea for an aggregate we are missing!

The analyticsCore/analytics\_core schema:

* **bucketid** – the “smart code” described above
* **blevel** – the bucket “level” breakout – for instance company\_id, state, ticker, country, soc code, naics code etc.
* **jobsactive** – the number of active jobs on the given date
* **jobscreated** – the number of created jobs on the given date
* **jobsremoved** – the number of removed jobs on the given date
* **durationActive** – the average length of time that the current active jobs have been posted for the aggregate
* **durationClosed** – the average length of time that the “removed” jobs were posted
* **jobsdate** – the date the analytic references. If you have 20 active jobs with jobsdate of 2020-01-01, this means “there were 20 active jobs on 2020-01-01”



You can find a list of all production aggregate core analytics here: <https://github.com/SmartMarketData/LinkUp_Enhanced_Documentation/blob/master/LU_Enhanced_Aggregate_Analytics_Codes.xlsx>

### Python Examples

With the analyticsCore files, you can get core job analytics for virtually any “bucket”, like tickers, company\_ids, etc. For instance, to get the point in time analytics for Deloitte & Touche (company\_id=’134’), you would pull the file from your local repository or S3 like this:

|  |
| --- |
| getblevel=’134  linkupfiledir=yourdrive+'LinkUp\\bucket\\'  bucketid='00011010000001'  getdir=linkupfiledir+bucketid+'\\analytics\_core\\'  analytics=pqc.ParquetDataset(str(getdir))#,filters=[(jobsdate', '=', '2020-08-28'),])  analytics=analytics.read()  analytics=analytics.to\_pandas()  coidpanel=analytics.loc[analytics.blevel==getblevel,:] |



To get Factset stock ticker for Accenture, you would replace ‘getblevel’ and ‘bucketid’ above.

|  |
| --- |
| getblevel=’ACN|NYS|US’  bucketid='00011030000001' |



Or for the state of California, again just replace the ‘getblevel’ and the ‘bucketid’.

|  |
| --- |
| getblevel=’CA’  bucketid='00011030000001' |

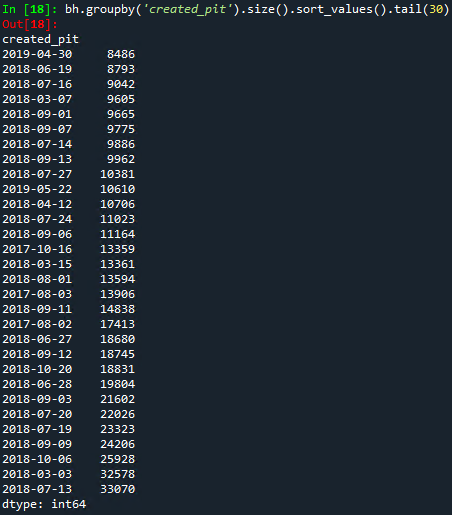


And we are adding more all the time.

analyticsEnhancedJob – **12.5GB**

This file is where we flag specific jobs , allowing you to exclude or slice them from the total panel of jobs. We currently have the following “analyticType” partitions:

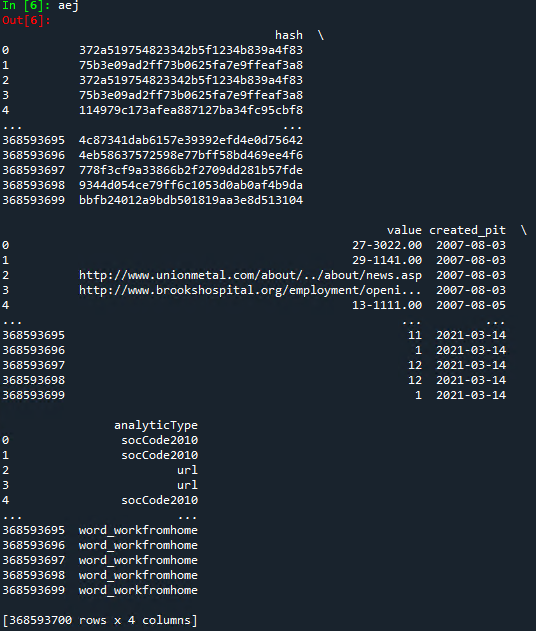
* **socCode2010** – the Bureau of Labor Statistics Standard Occupational Classification, see the “auxiliary/soccode/soc2010” parquet file for the helper table with the codes and definitions. This standard categorization from the Bureau of Labor Statistics is far superior with a 4 level schema and thousands of potential classifications vs. the old LinkUp “unique” category\_id which was very limited.
* **url** – the url used to scrape the job
* **word\_parttime** – hash will be flagged if part-time jobs – the flag is implied if the hash are in the panel. The flag is actually a “score” that we obtain from our proprietary job flagging model where we score the likelihood of the job being “part time” in this case. Higher “scores”, more likely a hit. Scores are absolute, and based on actually word matches. Hash that don’t appear in the panel had a zero score. There are over 16M part time jobs in the data set as of Q1 2021.
* **word\_workfromhome** - hash will be flagged if work from home or remote jobs – the flag is implied if the hash are in the panel. The flag is actually a “score” that we obtain from our proprietary job flagging model where we score the likelihood of the job being “wfh/remote” in this case. Higher “scores”, more likely a hit. Scores are absolute, and based on actually word matches. Hash that don’t appear in the panel had a zero score. There are almost 800K work from home/remote jobs as of Q1 2021.
* **badhash** – here we flag hash that were once in production, and have since been removed from the data set. Hash can be bad for many reasons, but in general they are found to be duplicates or scrape mistakes from runaway scrape processes (caught in a loop) etc… There are roughly 2M bad hash in the dataset currently. The vast majority were removed in a cleanup in 2018. Here is the distribution of cleanups:



* **reposts\_seo** – here we flag jobs that appear to have been “deleted”, then “reposted” in back to back scrapes. Companies do this to game the Search Engine Optimization (SEO), and keep their jobs “fresh” so they appear at the top of searches. We use this for further normalization analytics, smoothing out “false” deleted jobs… This helps smooth some of the noisiness that appears in the point in time analytics sets.

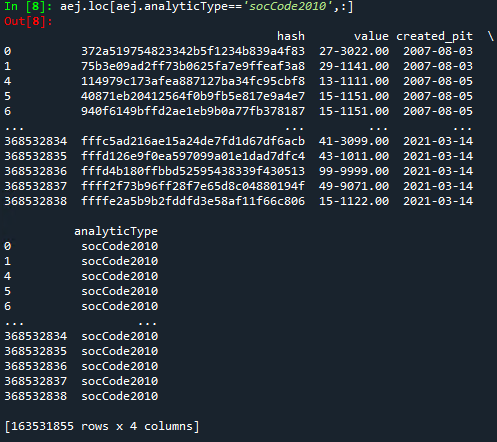
To access the file, open the parquet:



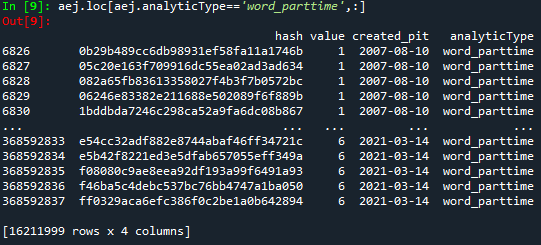
This brings the entire file into a dataframe: 

Which you can slice directly by “analyticType”, some examples:

**socCode2010**

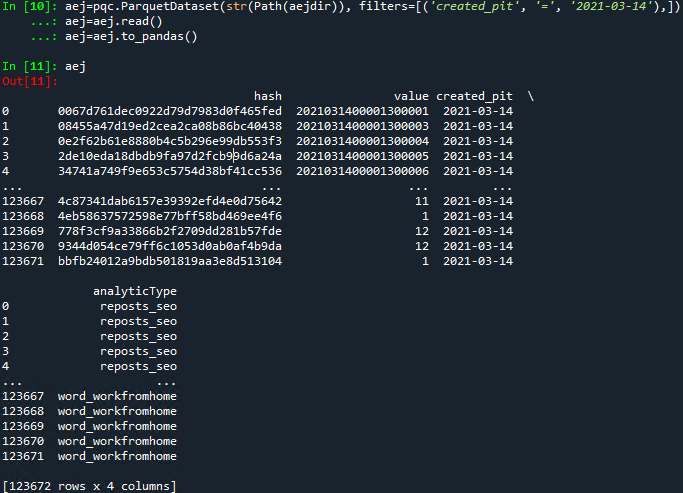


**word\_parttime**

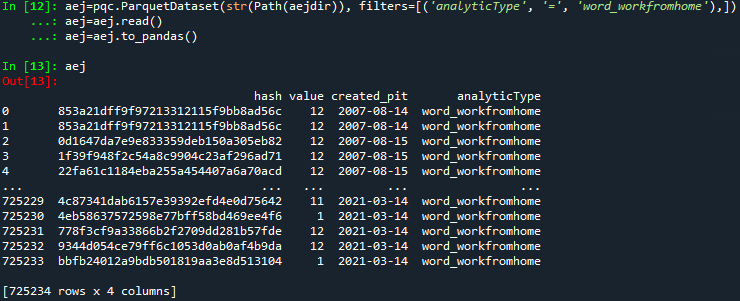


Or you can “filter” in the parquet call directly by “analyticType”, or by the “created\_pit” partition (the day the job was created), some examples:

**created\_pit partition**



**Or just partition out the “word\_workfromhome”**



File Glossary & Delivery Timing

* Glossary <https://github.com/SmartMarketData/LinkUp_Enhanced_Documentation/blob/master/LU_Enhanced_Glossary.xlsx>
* All files are in parquet format. Most are elegantly partitioned for selective access to slices of each file.
* AWS S3 – with sync and Glue capability
* Timing
  + scrapelog ~9:30PM EDT/~8:30PM EST (1:30AM UTC)
  + jobs\_base ~10:30PM EDT/~9:30PM EST (2:30AM UTC)
  + jobs\_log ~10:30PM EDT/~9:30PM EST (2:30AM UTC)
  + analyticsEnhancedJob ~9:45PM EDT/~8:45PM EST (1:45AM UTC)
    - analyticType=socCode2010 - ~9:45PM EDT/~8:45PM EST (1:45AM UTC)
    - analyticType=url - ~9:45PM EDT/~8:45PM EST (1:45AM UTC)
  + descriptions ~11:15PM EDT/~10:15PM EST (3:15AM UTC)
  + reference ~11:15PM EDT/~10:15PM EST (3:15AM UTC)
  + analyticsCore ~12:00AM EDT/~11:00PM EST (4:00AM UTC)
  + analyticsEnhancedJob ~12:00AM EDT/~11:00PM EST (4:00AM UTC)
    - analyticType=word\_parttime - ~12:00AM EDT/~11:00PM EST (4:00AM UTC)
    - analyticType=word\_workfromhome - ~12:15AM EDT/~11:15PM EST (4:15AM UTC)
    - analyticType=reposts\_seo - ~1:45AM EDT/~12:45AM EST (5:45AM UTC)
* The *reference* and *scrapelog* files are replaced each day. We would recommend looking at the modified time of the posted data, and if changed then reload the entire file. You can also just pick a random partition and use that as the check (i.e. action\_date=2018-01-01 or reftype=stock\_ticker\_primary for *scrapelog* and *reference* respectively). Another strategy is to just reload the entire file every day without a check. The auxiliary is just that, we will periodically place helper tables/files in there. The files will rarely change.
* The *reference* file has a “thrudate” field which is effectively the point in time that we added or changed the record.
* All of the raw jobs data and analytics/aggregate files are partitioned point in time. You can just pull the missing or latest partitions daily. We would notify you if we have need for a complete “reload” which would only occur if we found a material issue with the data or analytics.

# File Sizes

As of 2021-08-09, here are the full parquet file size footprints, as well as the daily partitions/files sizes.

Text

Description automatically generated with medium confidence

# AWS S3 access

We can provide access to all of the LinkUp jobs data and analytics through AWS S3.

The preferred method uses the best practices “role” model described below. As an alternative, we can create a secure user which can access the bucket with a unique Key/Secure Key, or a userid and password model if you would like to access via the AWS console, or a third party client such as FileZilla or CyberDuck.

## Details for implementing the “role” model

You give us an aws:arn. Typically this takes the form of your unique s3 bucket id. For instance:

**arn:aws:iam::123456789123:root**

with a best practice where you would instead assign to a specific “user”, something like:

**arn:aws:iam::123456789123:user/yourfund\_uploader**

We create the role then, which would look something like this:

**arn:aws:iam::389755834573:role/S3-SMD\_LinkUp-yourfund**

Once we give you the “role”(s) we create for your bucket(s), you would:

## AWS IAM Policy to Grant Access to LinkUp Data Files in S3

Once you have your organization's role name, you will need one of your AWS administrators to create a policy that will allow users or roles to assume the SMD provided role.

To create the policy from the AWS console:

1. From the Services menu select IAM.

2. From the left menu under Access Management select Policies.

3. Click the Create Policy button.

4. For Service search for and select STS.

5. For Actions expand the Write section under Access Level.

6. Check the box for AssumeRole.

7. Under resources ensure that Specific is selected.

8. Then for Role click Add ARN to restrict access.

9. For Specify ARN for Role enter the role provided to your organization by SMD

10. Click the Add button.

11. Click the Review policy button.

12. Give the policy a name. Your organization might have a preferred way to name policies. If you don’t, a possible name could be STS-assume\_SMD\_LU\_Data\_S3\_access\_role\_policy.

13. Add a policy description if you want.

14. Click the Create Policy button.

Now that the policy is created, you can attach it to AWS IAM users, groups, or roles, and of course you can setup to sync etc.

## Attach the AWS IAM Policy to Grant Access to LinkUp Data Files in S3 to an IAM user

To attach the policy to a user:

1. From the left menu under Access Management select Users.

2. Click on the name of the user you want to attach the policy to.

3. Under the Permissions tab click the Add permissions button.

4. Click Attach Existing Policies Directly.

5. Check the box next to the policy you created earlier.

6. Click the Next: Review button.

7. Click the Add permission button.

## Attach the AWS IAM Policy to Grant Access to LinkUp Data Files in S3 to an IAM group

To attach the policy to a group:

1. From the left menu under Access Management select Groups.

2. Click on the name of the group you want to attach the policy to.

3. Under the Permissions tab click the Attach Policy button.

4. Check the box next to the policy you created earlier.

5. Click the Attach Policy button.

## Attach the AWS IAM Policy to Grant Access to LinkUp Data Files in S3 to an IAM role

To attach the policy to a group:

1. From the left menu under Access Management select Roles.

2. Click on the name of the role you want to attach the policy to.

3. Under the Permissions tab click the Attach Policy button.

4. Check the box next to the policy you created earlier.

5. Click the Attach Policy button.

# FAQs

1. **How do I create my own analytics?**

You can create the analytics using the jobs\_log. At its base, if you just want a “Macro” look, you would query the addremoveflags 1=created\_pit, and -1=deleted. This gives you the bookends that each jobid (job) was active. There are any number of ways to finish the job based on where you hold your table, and the programming language you use. In python, you could unstack the table so you had a “jobid”, “created\_pit”,”deleted” in individual rows for each jobid. You could then loop or all at once count/sum/size the jobs where your target date is >= the created\_pit, and < the deleted.

To do this for company\_id’s (a “blevel” described above), you would first join/merge the jobs\_log with the jobs\_base, attaching the appropriate company\_id to the jobs\_log. Of course you can do this for any of the job specific items in the jobs\_base, like country, city, state, zip…

If you want ticker, or other reference data, you can join the company\_id as above, then join/merge the “reference” table, and attach the appropriate ticker, or other identifier you would like (like NAICS, or TRBC(if permissioned), or SEDOL, or PermID… You name it!).

WOnce attached, you would do the same exercise as above, but group by the ‘blevel’ you want (i.e. company\_id,ticker, sedol, etc…)

But don’t forget! We have already done most of these for you in the analyticsCore described above. And if we havn’t put in production yet something you want, let us know!

1. **The old LinkUp data had a deleted date for each “hash”. Where is it?**

The deleted date is a part of the “jobs\_log”. Each “jobid”, when deleted, is assigned a “-1” on the addremoveflag.

The jobs\_log also contains the created\_pit date(the pit corrected version of the old “created”) (“1”), as well as each and every scrape (“0”) which aligns with every “last\_checked” point in time from the old data set. Here we give you a full record. In the old LinkUp, you lost that hash level scrape information every day.

To capture all of the deleted dates, you can:

1. In the jobs\_log, query only for addremoveflag=-1. You will then have the deleted dates for each jobid.
2. You can extract directly from the parquet, as the file is partitioned by not only the scrape\_date, but also the addremoveflag. So in python, this would get you just the deleted dates:

**deleteddates=pqc.ParquetDataset(str(Path({locationofjobs\_log.parquet})), filters=[('addremoveflag', '=', '-1'),])**

1. **What is “created\_pit” and where is the old “created\_date”?**

LUGC has a “created\_pit” (this is the actual point in time production “created” date. LU assigns the created at the actual time the job was scraped, all good. BUT, they then assign the “last\_checked” date when the company\_id level scrape ended – which determines the date the information is published to users. This creates situations where in the daily scrape cycle, the job was created on 1 day, but “checked” on another. This creates many problems PIT, and creates a scenario where historical analytics are “changed” from day to day. With LUGC, jobs and analytics are perfectly synced with the production date – the day a client would have actually seen the information. As of August 5, 2021, there are 7.8M jobids that the **created\_pit** is > the **created** timestamp “date”, or about 4.5% of the jobs.

The **created** “timestamp” field is in the jobs\_base file. It is not used directly in the analytics, the “created\_pit” is. Many times they are the same. Both of these are immutable, and therefore in the jobs\_base file. You can also find the “created\_pit” as the first 8 digits of the jobid for each job. You can also capture the created\_pit from the job\_log with the same process as above for “deleted”, by using 1 instead of -1.

1. **Why are the daily AnalyticsCore (aggregate analytics) file sizes different day to day?**

The file size of the aggregate files is a function of “how many” aggregate constitutuents were updated for the day, which is a direct function of the number of company\_ids that were scraped, which is a function of the companies in the scrape queue and how many were actually processed for the day, which is a further function of the size of the companies in the queue in terms of the number of active jobs – i.e. more companies means longer time to scrape that company.

So for the company\_id level aggregate (00011010000001), you will see a direct correlation with the number of distinct company\_ids in the jobs\_log for the day, as well as tangentially in the jobs\_base (the jobs created). Today for instance was 234KB, and was 227KB yesterday.

For ticker levels, you would have smaller files and larger possible daily % deltas in file size. As you can imagine, we may scrape more or less listed companies on a given day.

For state level aggregates, you would expect smaller changes, as most states would get some representation in the days scraped jobs…

There is no expected range per se. I would just look at the min and max file sizes of each aggregate’s partitions, and allow for some leeway. We can have days where we scrape many companies, and occasionally days when relatively few. If we scrape few (could be a system problem, or just a lot of big companies, or system maintenance on a weekend), then the jobs\_log will show fewer distinct company\_id’s. You can also use the “scrapelog” and see which companies were scraped, and again will be a barometer for the relative sizes of most aggregates.

1. **What is the relationship between SmartMarketData LLC and JobDig, Inc.?**

SmartMarketData LLC (SMD) partnered with JobDig, Inc. (LinkUp) in 2014, to take their raw job post data, productize it, and create analytics for Wall Street. We are a partner across all business processes including product development, operations, business development and client administration. In 2021, SMD added a direct distribution model to the partnership allowing clients to receive enhanced product, analytics, and services directly from SMD.

1. **Why the change from .csv files to .parquet?**

Parquet files offer significant operational efficiencies over csv files. Please see these links for a nice explanation of the differences, the benefits, and what parquet files are.

[What is Apache Parquet? (databricks.com)](https://databricks.com/glossary/what-is-parquet)

[Apache Parquet](http://parquet.apache.org/documentation/latest/)

[Parquet Files - Spark 3.1.1 Documentation (apache.org)](https://spark.apache.org/docs/latest/sql-data-sources-parquet.html)

1. **What are “Reposts”, and “Repost Upward Drift”?**

Historically, on 2016-01-23, we saw the raw jobs data for 2010-01-01 as it was on 2016-01-23. We had no record backwards of what you would have seen “point in time”. While differences are fairly insignificant, there is a systemic upward “drift” in job counts over time, based on the analytics and data capture methodology, due to companies “reposting” jobs they had previously removed for a period. For instance, company XYZ may have 10 jobs active on 2010-01-01. On the next “scrape” on 2010-01-02 they have 9 jobs, having removed 1. There are 10 scrape days where no changes occurred, so on 2010-01-12 we still show 9 jobs, and the time series for the period would show 10, then 9,9,9...9. Now company XYZ adds back the EXACT SAME JOB (based on the unique URL) on 2010-01-13. For 2010-01-13 we will show 10 jobs again. *BUT*… the time series will now adjust back that that job existed for ALL the days, so it would look like 10,then 10,10,10…10. A “drift” updward of 1 job.

Again, NOT a PIT representation of what we knew on a specific day. We fix this nuance and many other things in R2E.

We eliminated the repost issue by capturing when companies remove jobs, and if they add back, creating a “new” job – that still references the original “hash”. This is accomplished with the jobid, and the parenthash in the jobs\_base\_file. In doing this, we eliminate the “drift” that occurs non-PIT. We also eliminate the problem of having a deleted date be NULL, then have a date, then suddenly be NULL again. Our goal is to capture when companies have jobs actually posted. This modified look gives you just that.

## PII – Personally Identifiable Information

The Job description raw data may contain PII (like job recruiter name and contact info) We do not attempt to remove them as it is all publicly available information, The percentage is insignificant though, <1%. As an alternative, we can remove Job Descriptions from the feed if this is an issue, and can also provide derived analytics on the Descriptions as an option. 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. As of 3/22/2021 there are roughly 6,310,775 email addresses that are now cleansed in the data set. Note there still may be a very small percentage of formal names still in the descriptions, that are again already publicly distributed via the internet.

## What is the “main company\_id” for a ticker?

There can be a many to one relationship between company\_id’s and tickers. This can happen because we add/change a scraper, the company changes the base url for a career portal, the company\_id represents a different corporate entity/child of the ticker, i.e. a subsidiary, affiliate, division, you name it. We assign with Factset, the Factset\_id. If you have Factset, you would look to see if that Factset\_id (which also has a many to one relationship to a ticker) is a “parent” or “child”. If a parent, then we would consider that the “main” company\_id. Factset does not allow us to distribute the “parent” Factset\_id, we would need to work with them to get it for you if interested. But you should not need to really worry about this. For instance, Berkshire Hathaway actually does not have a career portal, but has some 40+ entities that roll up to it (i.e. Geico insurance). So there would be no “main company\_id”.

1. **Why are there some instances of a “parenthash” falling under more than one company\_id?**

As of 2020-08-05 there are roughly 71,000 total parenthash out of close to 180M, which have the distinction of being under 2 or more company\_id. While each parenthash is unique, and based on a unique URL, there are occasions that related companies that we scrape, may have used the same URL as the source for the job. So company XYZ and company ABC may be related and post job 123, which has a single URL, an each individual company career portal. The total number of instances is insignificant, but does happen. There is no real way for us to deal with this because we do not want to be in the business of deciding which one should be deleted. We do deal with this effectively at the jobid level though.

1. **Why would the “created” timestamp date be GREATER THAN the “created\_pit” date?**

This should not happen conceptually. It is related to the multiple company\_id to parenthash issue in the previous FAQ. As of 2021-08-05 there are 182K jobids that have this issue, which is inline with roughly 2X+ the dupe company\_id to parenthash issue. Again, this is insignificant at 0.1% of jobs. Even more insignificant when you factor that this impacts only ~4,700 company\_ids. We would recommend ignoring. Worst case, they are related companies, and when aggregated at the ticker/sedol etc. level, they double count. But the impact will be negligible. From the scrape\_team there was a “bug” which when filtering company\_id’s was causing “new” create dates if/when a job hash switched from one company\_id to another. This bug was fixed in June of 2020. You should not see these in general go forward.

1. **What is your International vs. USA mix job post mix?**

LinkUp Gold Copy has job posts from 240 countries worldwide. As of 2021-08-05 here are the top 20 countries and their job counts and the first job scraped “date”:

A picture containing graphical user interface

Description automatically generated